

Statistical Add-Ins for Excel

User Guide



Statistical Add-Ins for Excel, User Guide

© The Numerical Algorithms Group Limited, 2000

All rights reserved. No part of this manual may be reproduced, transcribed, stored in a retrieval system, translated into any language or computer language or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the copyright owner.

The copyright owner gives no warranties and makes no representations about the contents of this manual and specifically disclaims any implied warranties or merchantability or fitness for any purpose.

The copyright owner reserves the right to revise this manual and to make changes from time to time in its contents without notifying any person of such revisions or changes.

Second Edition

**ISBN 1-85206-191-X
NP3521**

NAG is a registered trademark of:

The Numerical Algorithms Group Limited
The Numerical Algorithms Group Inc
The Numerical Algorithms Group (Deutschland) GmbH
Nihon Numerical Algorithms Group KK

All other trademarks are acknowledged.

Contents

1. INTRODUCTION.....	1
2. INSTALLATION.....	2
3. USING THE ADD-INS.....	3
3.1. LOADING AN ADD-IN	3
3.2. SIMPLE USAGE	3
3.3. PARAMETER INPUT	4
3.4. ARRAY FUNCTIONS	4
3.5. ORGANISING THE OUTPUT	5
3.6. GETTING ONLINE HELP	6
3.7. ERROR MESSAGES	6
3.8. CHANGES FROM RELEASE 1.0.....	7
3.9. FURTHER TECHNICAL INFORMATION	8
4. ADD-INS IN BOTH SETS.....	9
4.1. NAG ADD-IN UTILITIES.....	9
4.2. NAGEXTRACT	9
4.2.1. <i>Introduction</i>	9
4.2.2. <i>Initialisation</i>	9
4.2.3. <i>Relocating the Results</i>	9
4.2.4. <i>Setting the Options</i>	10
5. NAG BOOKS SET 1 (MODELLING AND MULTIVARIATE METHODS).....	13
5.1. INTRODUCTION.....	13
5.2. BOOK 0.....	13
5.2.1. <i>Introduction</i>	13
5.2.2. <i>Distribution Functions</i>	13
5.2.2.1. Introduction	13
5.2.2.2. NORMAL_PROB (Set 1).....	14
5.2.2.3. NORMAL_PPT (Set 1)	15
5.2.2.4. MULTI_NORMAL_PROB (Set 1)	15
5.2.2.5. T_PROB (Set 1).....	16
5.2.2.6. T_PPT (Set 1)	16
5.2.2.7. CHI_PROB (Set 1)	16
5.2.2.8. CHI_PPT (Set 1).....	17
5.2.2.9. F_PROB (Set 1).....	17
5.2.2.10. F_PPT (Set 1)	17
5.2.2.11. BETA_PROB (Set 1).....	18
5.2.2.12. BETA_PPT (Set 1)	18
5.2.2.13. GAMMA_PROB (Set 1)	19
5.2.2.14. GAMMA_PPT (Set 1).....	19
5.2.3. <i>Summary Statistics</i>	20
5.2.3.1. Introduction	20
5.2.3.2. SUMMARY_STATS (Set 1).....	20
5.2.3.3. S5PT_SUMMARY (Set 1).....	21
5.2.3.4. TWO_WAY_TABLE (Set 1).....	21

5.3.	Book 1	22
5.3.1.	<i>Introduction</i>	22
5.3.2.	<i>Regression and Correlation</i>	22
5.3.2.1.	Introduction	22
5.3.2.2.	CORREL_MAT (Set 1)	25
5.3.2.3.	PARTIAL_CORREL_MAT (Set 1)	25
5.3.2.4.	MULT_LIN_REG (Set 1)	26
5.3.2.5.	MM_MLR (Set 1)	27
5.3.2.6.	DURBIN_WATSON (Set 1)	29
5.3.2.7.	STAND_RESIDS (Set 1)	29
5.3.2.8.	DUMMY_VARS (Set 1)	30
5.3.3.	<i>Time-Series</i>	31
5.3.3.1.	Introduction	31
5.3.3.2.	TIME_SERIES_DIFF (Set 1)	34
5.3.3.3.	ACF (Set 1)	34
5.3.3.4.	PACF (Set 1)	35
5.3.3.5.	ARIMA_APPROX_FIT (Set 1)	35
5.3.3.6.	ARIMA_FIT (Set 1)	36
5.3.3.7.	ARIMA_FORECAST (Set 1)	38
5.3.3.8.	TRNS_FUNC_FORECAST (Set 1)	38
5.3.3.9.	SPECTRAL (Set 1)	40
5.4.	BOOK 2	41
5.4.1.	<i>Introduction</i>	41
5.4.2.	<i>Analysis of Experiments</i>	41
5.4.2.1.	Introduction	41
5.4.2.2.	ANOVA_BLOCK (Set 1)	46
5.4.2.3.	ANOVA_ROW_COLUMN (Set 1)	47
5.4.2.4.	ANOVA_CONTRASTS (Set 1)	48
5.4.2.5.	MULT_COMPN (Set 1)	49
5.4.2.6.	ANOVA_FACTORIAL (Set 1)	50
5.4.3.	<i>Generalized Linear Models</i>	51
5.4.3.1.	Introduction	51
5.4.3.2.	NORMAL_GLM (Set 1)	53
5.4.3.3.	BINOMIAL_GLM (Set 1)	55
5.4.3.4.	POISSON_GLM (Set 1)	58
5.4.3.5.	GAMMA_GLM (Set 1)	60
5.4.3.6.	MM_GLM (Set 1)	62
5.5.	BOOK 3	63
5.5.1.	<i>Introduction</i>	63
5.5.2.	<i>Variable-Directed Methods</i>	63
5.5.2.1.	Introduction	63
5.5.2.2.	PRIN_COMP (Set 1)	65
5.5.2.3.	FACTOR (Set 1)	66
5.5.2.4.	FACTOR_SCORE (Set 1)	67
5.5.2.5.	ORTHOG_ROTATION (Set 1)	67
5.5.3.	<i>Individual-Directed Methods</i>	68
5.5.3.1.	Introduction	68
5.5.3.2.	DIST_MAT (Set 1)	73
5.5.3.3.	CLUSTER (Set 1)	73
5.5.3.4.	GROUPS_FROM_CLUSTER (Set 1)	75
5.5.3.5.	DISCRIM_ALLOC (Set 1)	75
5.5.3.6.	DISCRIM_DIST (Set 1)	76
5.5.3.7.	DISCRIM_TEST (Set 1)	77
5.5.3.8.	MDS (Set 1)	78
5.5.3.9.	METRIC_SCAL (Set 1)	79

6. NAG BOOKS SET 2 (NONPARAMETRIC METHODS)	81
6.1. INTRODUCTION	81
6.2. BOOK 4	81
6.2.1. <i>Introduction to Nonparametric Methods</i>	81
6.2.2. <i>Index of Nonparametric Methods</i>	83
6.2.2.1. NONPARAMETRIC_INDEX (Set 2)	83
6.2.3. <i>One Sample</i>	84
6.2.3.1. BINOMIAL_TEST (Set 2)	84
6.2.3.2. CHI_SQ_1 (Set 2)	85
6.2.3.3. COX_STUART (Set 2)	85
6.2.3.4. RUNS_TEST (Set 2)	86
6.2.3.5. WILCOXON (Set 2)	86
6.2.3.6. ONE_SAMPLE_CI (Set 2)	87
6.2.3.7. PROPORTION_CI (Set 2)	88
6.2.4. <i>Matched Pairs Samples</i>	88
6.2.4.1. MCNEMAR (Set 2)	88
6.2.4.2. SIGN_PAIRS (Set 2)	89
6.2.4.3. WILCOXON_PAIRS (Set 2)	89
6.2.5. <i>Two Samples</i>	90
6.2.5.1. CHI_SQ2 (Set 2)	90
6.2.5.2. KOLMOGOROV_2 (Set 2)	91
6.2.5.3. MANN_WHITNEY (Set 2)	91
6.2.5.4. MEDIAN_TEST (Set 2)	92
6.2.5.5. TWO_SAMPLE_CI (Set 2)	92
6.2.6. <i>K Samples</i>	93
6.2.6.1. CHI_SQ_K (Set 2)	93
6.2.6.2. COCHRAN_Q (Set 2)	94
6.2.6.3. FRIEDMAN (Set 2)	94
6.2.6.4. KRUSKAL_WALLIS (Set 2)	94
6.2.7. <i>Correlation and Association</i>	95
6.2.7.1. CONTINGENCY_C (Set 2)	95
6.2.7.2. KENDALL_CC (Set 2)	95
6.2.7.3. KENDALL_TAU (Set 2)	96
6.2.7.4. SPEARMAN_RHO (Set 2)	96
7. TROUBLESHOOTING GUIDE	97
7.1. INTRODUCTION	97
7.2. QUESTIONS	97
8. APPENDICES	100
8.1. GLOSSARY	100
8.2. ALGORITHMIC ASSOCIATION	101
8.2.1. <i>Add-In / NAG Subroutine Correspondence</i>	101
8.2.2. <i>NAG Subroutine / Add-In Correspondence</i>	102
8.3. REGISTRATION	105
9. CONTACTING NAG	107
10. REFERENCES	108

1. Introduction

This manual describes the NAG Statistical Add-Ins for Excel, their installation and their usage.

The underlying algorithms are based upon routines from the NAG Fortran Library and detailed descriptions of these algorithms are contained in the NAG Fortran Library manuals. These manuals may be purchased from NAG. Alternatively, useful information may be obtained from the NAG web sites.

The Add-Ins have been prepared under Excel version 7 on a PC running Microsoft Windows 95, and tested under Excel 97 and Office 2000 running Windows NT, Windows 95 and Windows 98.

Both the 'NAG Statistical Add-Ins for Excel: Modelling and Multivariate Methods' and the 'NAG Statistical Add-Ins for Excel: Nonparametric Methods' are documented here. The Add-Ins are referred to as 'Set 1' and 'Set 2' respectively in the individual function documents. If you have purchased just one of these products then you will not be able to use functions from the other. NAG will of course be pleased to accept an order for the other set.

The items described as belonging to Book 0, Book 1, Book 2 and Book 3 together form the 'NAG Statistical Add-Ins for Excel: Modelling and Multivariate Methods' product. Items belonging to Book 4 constitute the 'NAG Statistical Add-Ins for Excel: Nonparametric Methods' product.

NAG is always anxious to receive feed-back from its users, so please contact us if you have any suggested improvements or if there is any extra functionality you would like.

Please register with us, either via the Web (www.nag.co.uk or www.nag.com) or by returning the form printed in the appendix of this manual. (This form may be printed from the help file if you do not want to take pages from your printed manual.) Once you are registered we will be able to keep you informed of any updates to this product and other enhancements from NAG.

2. Installation

You should have a CD containing the Add-Ins material or a file downloaded from the web.

If you have a CD then, under normal circumstances, the CD will autorun. If the autorun facility is disabled, double click on the setup.exe file on the CD from Windows Explorer. Follow the prompts on the screen to complete the installation process.

If you have downloaded a file from the web, then double click on the setup.exe file from Windows Explorer and follow the prompts on the screen to complete the installation process.

The Add-Ins will not work if the installed materials are simply copied from one machine to another. If the Add-Ins are required on additional machines then check that the licence with NAG covers the additional machine and re-install from the CD or file provided.

The DLL is installed into a directory that **must** be on the user's PATH and the installation procedure asks for permission to edit the autoexec.bat file to ensure this. We recommend that you allow this.

To verify that the installation process has completed correctly, re-boot the system. Open Excel with a new workbook. From the 'Tools\Add-Ins' menu, load the NAG Add-Ins by ensuring that the appropriate NAG check boxes contain a 'tick'. If not, use the left button of the mouse to click on the appropriate check box. If the NAG Add-Ins load without any warnings then the system information has been correctly installed.

As a simple run-time test, if you have the NAG Statistical Add-Ins for Excel: Modelling and Multivariate Methods product, then select a cell on a worksheet and enter the formula =CHI_PPT(0.5,1). If the cell evaluates to 0.454936... then all is well and the DLL is indeed being correctly accessed. If you have the companion product for nonparametric methods then PROPORTION_CI is a suitable test function. Select a cell on a worksheet and enter the formula =PROPORTION_CI(2,4). The cell will evaluate to <List of 5> if the product has been successfully installed. The meaning of this output is explained later in the section on array functions.

3. Using the Add-Ins

3.1. *Loading an Add-In*

In order to use the NAG Add-Ins, they must first be loaded into Excel. To do this, use the Tools\Add-Ins menu and select the 'Books' required. The Add-In, 'NAG Add-In Utilities', or NAG_Library.xla, is referenced by the other NAG books and theoretically should not need to be explicitly loaded. In practice we have often found it necessary for the user to do this. You are advised to experiment with your system to determine whether it is necessary in your particular case.

If all has gone well, you should see a 'Nag' item appear on the menu bar. If not, then you will have obtained a message box indicating that either an error has occurred in the installation process or the system has been corrupted. You should note that installation must be performed via the NAG setup.exe routine. It does not suffice to copy files from one machine to another.

Please note that a complete set of appropriate examples are provided in the Excel workbooks. These are located in the 'Examples' subdirectory of the Add-Ins directory and on the CD or other media. After you have read other sections in this chapter, you might find it helpful to work through the examples to familiarize yourself with the use of the Add-Ins. In particular we recommend that you study closely the section on Array Functions, since the Add-Ins make extensive use of this Excel facility.

3.2. *Simple Usage*

From the 'Nag' menu choose 'About'. A message box appears reflecting the personal details you gave on installation.

If you have the NAG Statistical Add-Ins for Excel: Modelling and Multivariate Methods product (Set 1), select a cell on a worksheet and enter the formula =CHI_PPT(0.5,1). A value of 0.454936... appears in the cell.

We can also call this NAG function from the **function wizard**. To do this, select a different cell on the worksheet and from the Toolbar click on the function wizard button and select the 'Statistical' Category. Under the 'Function name' panel select CHI_PPT. You will see that this requires two parameters. There is also a brief statement of the routine's functionality. In this case, CHI_PPT returns the deviate associated with the given lower tail probability of the chi-squared distribution with real degrees of freedom. Click the 'Next' button. On the new dialog box, select the first parameter and type 0.5. Do **not** press enter at this stage. Select the second parameter and enter the value 1. Either click on the 'Finish' button or press 'Enter'.

We can refine this process somewhat by entering data in the cells. Select a cell and type in the value 0.5. Now select another cell and enter the value 1. Select a different cell to contain the final result and, as before, enter the function wizard for CHI_PPT. Click the dialog box for the first parameter and now, with the mouse, select the cell containing the value 0.5. The cell address (e.g. A4) will appear in the dialog box as the first parameter. Similarly, click the cell containing 1 to specify the second parameter. Press 'Enter' or click the 'Finish' button. Again the value 0.454936 should appear in the output cell.

If we now change the value 0.5 to say 0.6 in the data cell, then with normal settings of Excel that allow automatic updating of cell values the result from the last computation will automatically be updated to 0.708326... to reflect a revised call of CHI_PPT with parameters 0.6 and 1. This is a very powerful feature of Excel functions that we sought to preserve with more complicated routines within the NAG Statistical Add-In suite.

If you only have the NAG Statistical Add-Ins for Excel: Nonparametric Methods product (Set 2) then select a cell on a worksheet and enter the formula =PROPORTION_CI(5,10). Text, <List of 5>, appears in the cell. Section 3.4 will explain the significance of this; suffice to say that this text indicates that the Add-In is functioning correctly. As in the example described above, PROPORTION_CI may be called from the function wizard and may take as its arguments data entered into cell locations in the workbook. For example cell A1 may contain the value 5 and cell B1 may contain the value 10. The call =PROPORTION_CI(A1,B1) is then equivalent to the previous usage.

See 'Array Functions' for more information

3.3. *Parameter Input*

The parameters of the NAG functions may require input of the following types:

- Single real number
- Single integer
- Single string
- A list or one-dimensional array of cells
- Two-dimensional array of cells, i.e., rows and columns

Single numbers or characters can be entered directly into the function wizard or by reference to cells. In the case of the string input only the first character is considered and it is not case sensitive, so

Y
yes
Ya

will be treated identically.

When a list of names or a single variable is required, this can be input either as a column selection of cells or as a row selection of cells. Alternatively, it can be entered into the function wizard as a list between curly brackets, for example, {1,0,1,0}.

Where two-dimensional arrays are expected, even if in a particular case there is only one column, they must be entered by reference to the cells. Variables are always assumed to be stored column-wise.

3.4. *Array Functions*

Many statistical calculations return more than a single value as their result. Sometimes they might return a number of scalar values, as in the case of S5PT_SUMMARY, whilst at other times they might return a number of arrays, of different shapes and sizes, together with scalar output. In order to preserve the automatic updating property of functions, NAG elected to implement the Add-Ins as **array functions**.

Microsoft provide a range of these in their own Add-Ins and we have followed their conventions.

To illustrate the use of Excel array functions, type in a 2×2 array in a block of 2×2 adjacent cells within a worksheet. For example, let us say D1:E2 contains the values. Now select another block of 2×2 cells and type in another set of values to represent a 2×2 array. We are going to multiply these two arrays together using the standard Microsoft array function MMULT. Since we know that the product of these matrices is itself a 2×2 matrix, we select a block of 4 cells to contain the result. Now in the **function wizard** select MMULT from the 'Math & Trig' category. Select one of the arrays, say D1:E2, with the mouse to satisfy the first parameter,

array1, and select the other array to satisfy the second parameter, array2. Press the 'Finish' button or simply press return. You will only have a single scalar value returned in the (1,1) position at this stage. Now click on the formula bar and press 'Enter' whilst holding down both the 'Control' and 'Shift' keys at the same time (CONTROL+SHIFT+ENTER). You should now see the result of the matrix multiplication appearing in the four cells selected for this purpose. Look also at the formula bar. The formula is enclosed in curly braces, to denote that it is an array formula. **You cannot achieve the same effect by typing in the curly braces explicitly; the CONTROL+SHIFT+ENTER mechanism is mandatory.**

If you have the NAG Statistical Add-Ins for Excel: Modelling and Multivariate Methods product then it is possible to use S5PT_SUMMARY in a similar fashion. This returns a set of useful numbers (the minimum, the lower hinge, the median, the upper hinge and the maximum) relating to an array of data. Firstly construct some initial data, for example the values 1,2,3,4,5,6 in contiguous cells of the worksheet, A9:A14. Select a column of 11 contiguous cells to contain the output from S5PT_SUMMARY — we will see why 11 cells and not 5 are required shortly. Now use the function wizard to enter the cells A9:A14 as the only argument to S5PT_SUMMARY. Press 'Enter' or click on the 'Finish' button. As with the matrix multiplication example above, we see that only the first cell of the selected cells contains any output; we have to enter the formula as an array formula. So we click on the formula bar and press CONTROL+SHIFT+ENTER. Now we see all 11 cells filled with useful information and note that, just as before, the array formula appears in the formula bar surrounded by curly brackets.

At the head of the column we see < List of 11>. This is very useful information which we will utilise later. It tells us that to store all the output from the Add-In we will require at least 11 cells. The second entry is [Minimum]. This tells us that the value underneath, in the third location, contains the minimum of the input data. The '[' is a special delimiter which is detected by a NAG utility (NAGExtract) which is used in more complicated applications and which will be described later. Had the Add-In failed, an error message would have been produced in the second cell of the selected area.

Our example with S5PT_SUMMARY presupposed that we knew that we had to select 11 cells to store the result. If we did not know that, then we observe that the first cell tells us this, so if we were to call S5PT_SUMMARY as though it were a normal scalar function, then this key information would be returned. Specifically, select a single cell on the worksheet (but make sure that there are vacant cells beneath it). Use the function wizard to select cells A9:A14, as before, as the argument to S5PT_SUMMARY and press the 'Enter' key or 'Finish' button. The selected cell now contains the <List of 11> information. Click on this cell to select it and drag down a further additional 10 cells so that 11 cells are now selected. Make sure you do not accidentally use the **fill handle** to drag the cell; simply click on the centre of the cell and drag downwards. Now click on the formula bar, where =S5PT_SUMMARY(A9:A14) is still displayed, and press CONTROL+SHIFT+ENTER. The selected cells now contain the required output.

To demonstrate the usefulness of array functions, change a value in one of the cells A9:A14. Notice that the output values change automatically.

3.5. *Organising the Output*

We might be very pleased with the results of our computation, but not require the output in contiguous locations. In fact we might wish to place the 5 values in different cells on different worksheets. We can use the Excel Paste-Link facility to do this. Select one cell of the output. Let us suppose it is the 1, corresponding to the minimum of the set of values. Copy this into the clipboard either by using the Edit\Copy menu or the copy button on the toolbar. Now select the new destination cell. From the Edit\Paste Special menu, click the Paste Link button. The value, 1, of the original cell appears in the new destination cell. If however the input data changes, so that 1 is no longer the minimum value produced by S5PT_SUMMARY, then both the original and

the new destination cells change automatically. This is because the new destination cell contains a reference to the original output cell, not just a copy of the original value.

We will now use NAGExtract, a macro provided by NAG in a separate Add-In. Ensure that the NAGExtract Add-In is loaded and click on the NAGExtract menu from the 'Nag' menu on the menu bar. We need to tell NAGExtract where the raw output cells are. Select the actual cells with the mouse on the worksheet. The entry may read something like \$F\$10:\$F\$20 if these cells contain the raw output. Now click the 'Search' button on the NAGExtract dialog box. The 'Available Data' area will now contain 5 entries: Minimum, Lower Hinge, Median, Upper Hinge and Maximum. This tells us that the items we can handle are five scalar elements, so in this instance the values we can relocate consist of 5 real numbers. Specify a new destination cell by deleting any default entry in the 'Destination Cell' box in the NAGExtract dialog box and selecting a fresh cell in the workbook. By default, 'Minimum' is selected as the data item to relocate. If desired, select a different entity with the mouse in the 'Available Data' list. Click on 'Finish' in the dialog box and you should see the required output value paste-linked to the new location. The new version of NAGExtract provided with both these sets of Add-Ins now, by default, provides a caption to the data items paste-linked. This facility may be turned off, using the 'Options' button of NAGExtract described later.

If we had chosen a different function, then some of the output might have consisted of array information. We would have used NAGExtract in exactly the same way, noting that when we specified the 'Destination Cell', it was necessary to specify just one location. By convention, this cell is deemed to be the top left-hand corner of any output array.

Of course, if all that NAGExtract did was simply to simulate the copy and Paste-Link facility of Excel, then it would be unnecessary to provide the macro. It is most useful when the output from a NAG routine includes matrix or vector output. The underlying routines present these in different forms which need to be interpreted correctly by NAGExtract and expanded to fill the correct cells of the worksheet. In the simplest case, a list (which is output as part of a column vector) may represent a column vector or a row vector. The 'Options' button on the NAGExtract dialog box allows you to specify which is required. Similarly, array information may be stored by row or by column, or may represent a triangular matrix. Again the representation is specified using the 'Options' button.

3.6. **Getting Online Help**

Online help may be obtained at any time from the 'Nag' menu by clicking the NAGHelp menu item.

Additionally, help on a specific routine may be obtained by clicking the 'Help' button on the **function wizard**, once a particular routine has been selected.

3.7. **Error Messages**

If the input to a NAG Excel function is invalid, an error message will be returned, rather than any numerical output. If the error is one of value or shape rather than type, for example if the formula =CHI_PPT(2.0,1) was entered, the following message would be returned.

'Error: Probability >= 1'

Similarly, if in a regression model fitted by the function MULT_LIN_REG the Y and X arrays were of different lengths, the following error message would be returned.

'Error: y and x of different lengths'

If, however, the input were of the wrong type, for example if the formula =CHI_PPT(.5,A) were entered, then the Excel error message

'#VALUE!'

would be returned.

When parameter values for some of the complex functions are being entered, an error message may temporarily be displayed until the complete information is given. For example, with the function MULT_LIN_REG, if the Y array has been entered, the error message

'Error: y and x of different lengths'

may be displayed on the function wizard while the X array is being selected until the complete X array has been selected.

Most NAG error messages should be self-explanatory; however, two need further explanation.

The error message

'Catastrophic failure'

indicates that one of two types of error has occurred:

An error that should have been trapped and reported has for some reason not been detected. If you get such an error, first check your inputs to the function. It is likely that one of the parameters is incorrect. If not, check the NAG Web site to see if there is a known problem with the function. If it is not recorded, please contact NAG, preferably by email, to report the problem.

A trial version of the Add-Ins has been installed and this has now expired. Please contact NAG to obtain a fully licensed copy.

The error message

'Error: Computational failure'

indicates that there has been a failure in the underlying algorithm. Such an occurrence should be extremely rare. If it does occur then, if the function has a tolerance parameter, try making it larger. Alternatively, try using a simpler model. If the problem persists with a range of problems please check the NAG Web site to see if there is a known problem with the function. If it is not recorded, please contact NAG, preferably by email, to report the problem.

3.8. *Changes from Release 1.0*

Users of Release 1.0 of the NAG Statistical Add-Ins for Excel: Modelling and Multivariate Methods should note that while the basic functionality has not been changed between Release 1.0 and subsequent releases there has been a major rewrite of the interfaces taking into account comments from users. The interfaces of Release 1.0 are not compatible with the subsequent interfaces. We hope users find the new interfaces easier to use and with the improved Help system will quickly adapt to them. Any future releases of these Add-Ins should not involve such a major change to the interfaces.

An additional functionality is provided by the new MM functions that can specify a linear model by means of a model formula. There is also improved functionality in NAGExtract.

3.9. *Further Technical Information*

The User Guide should provide you with all the information required to run the NAG Statistical Add-Ins. If you want further information on the algorithms used to compute the statistics, please refer to the appropriate routine document from the NAG Fortran Library. Printed versions of the manual are available from NAG. Alternatively you may wish to look at the documentation available on the NAG Web sites: www.nag.co.uk and www.nag.com.

4. Add-Ins in Both Sets

4.1. *NAG Add-In Utilities*

It is not envisaged that you will need to call the NAG utilities directly under normal circumstances. Each of the NAG Add-Ins needs to perform various common actions, such as taking information from the worksheet and putting the values into appropriate Visual Basic arrays. These shared utilities are grouped together in the 'NAG Add-In Utilities' Add-In for greater efficiency.

Theoretically this Add-In should not require explicit loading, since it is **referenced** by each of the NAG Add-Ins that require it. We have found that it may be necessary to explicitly load the Add-In on some systems. **We recommend therefore that you experiment to see whether this is indeed the case for your own system.**

4.2. *NAGExtract*

4.2.1. Introduction

NAGExtract is a utility which may be invoked from the Nag menu on the menu bar. Its purpose is to ease the paste-link burden as you seek to reposition elements of the output to disparate positions. Whilst individual elements and possibly even simple vectors can more readily be repositioned using the Microsoft copy and paste-link facilities, the utility is especially useful if the output contains two-dimensional arrays, packed arrays or triangular arrays.

See 'Organising the Output' (Section 3.5) for a perspective of this utility. 'Initialisation' deals with getting information into NAGExtract, 'Re-locating the Results' describes the mechanism of doing a past-link using NAGExtract and 'Setting the Options' describes how to set the options for the different types of results obtained from NAG Add-Ins.

Please note that a complete set of appropriate examples are provided for each product in Excel workbook. These are installed in the 'Examples' subdirectory of the Add-Ins directory and on the CD or other distribution media. You might find it helpful to use them to practice the use of NAGExtract.

4.2.2. Initialisation

Before relocating the output, NAGExtract needs to be loaded with the data output from the NAG Add-Ins. Display the NAGExtract dialog box by selecting NAGExtract from the NAG menu. The 'Location of raw NAG results' edit box is empty and has the focus, i.e., this box is ready for you to put information into it. Do this by typing in the range of cells containing the output from the NAG Add-In, or alternatively select the range of cells using the mouse. Finally, press the 'Search' button. You should see a number of items appear in the 'Available data' window.

NAGExtract is now initialized, ready for use and the 'Apply' and 'End' buttons are enabled.

4.2.3. Relocating the Results

From the 'Available Data' box, select the data item that needs to be re-positioned. Then click on the 'Destination Cell' input box. The chances are that the cell location contained therein is not the one you require, so delete or select the current contents for overwriting. Next specify the new location of the output by typing in, or selecting with a mouse, a single cell location. With default settings this specifies the top left-hand corner of the output; if the output is a scalar and headings are not required (see below) then the selected cell is simply the location of the output.

For scalars and ordinary arrays, press the 'Apply' button. This paste-links the material to the specified location and leaves the NAG dialog box in place ready to re-position other elements of the output. To copy and close the dialog box press the 'End' button. To simply close the dialog box without copying press the 'Cancel' button.

The 'Options' button is used to tell NAGExtract how to unpack output that represents two-dimensional arrays, but which is output in some packed form. Select the appropriate option. The options also allow the user to specify transposition of vectors or to specify whether triangular matrices are to output as a list of elements or as a triangular array. See 'Setting the Options' for more detail.

The 'Help' button brings up help on the NAG Add-Ins.

An alternative approach is to use the Excel function INDEX. This allows you to pick out a single element of an array. So, for example, INDEX(PROPORTION_CI(5,10),3) would calculate the third element of the NAGExtract format list produced by the function PROPORTION_CI. Of course, INDEX may also be used on the NAGExtract format list itself, so for example INDEX(A1:A5,3) would copy the third element of the NAGExtract format list, stored in A1:A5.

4.2.4. Setting the Options

The Options Dialog Box

The Options dialog box contains a set of radio buttons and check-boxes which control the output of arrays and vectors. Once your selection has been made, click on OK to close the window. Changes to the default options are only temporary. The next time you call NAGExtract, or the options menu of NAGExtract, the options will return to the default settings.

Two-dimensional Array Handling Options

The default handling of two-dimensional arrays is to pack them in [row, column] style, or 'row order', the Fortran and Visual Basic convention. However, if you so wish, the two-dimensional arrays can either be transposed, with the 'Pack 2-D arrays in Column Order' radio button, or decomposed into a one-dimensional list, with the 'Pack 2-D arrays as a List' option.

Transpose Vectors (One-dimensional Arrays) Option

This option allows you to change the way NAGExtract outputs one-dimensional arrays from columns to rows. If the default is un-checked the vectors are output in a column.

Output Triangular Matrices Option

This option allows you to control whether triangular matrices are output as lists, or as triangles. The default is to output them as triangles.

Output Headings Option

This option controls whether or not the name of the entity (as specified in the output list) is also placed in the worksheet. If so it appears, in bold type, centred across the columns containing the entity. The default action is to output the headings.

OK Button

This button sets the options for the duration of the NAGExtract call or until you reset them.

Help Button

The Help button displays information on NAGExtract.

5. NAG Books Set 1 (Modelling and Multivariate Methods)

5.1. *Introduction*

There are 6 Add-Ins provided with the NAG Statistical Add-Ins for Excel: Modelling and Multivariate Methods product:

NAG Add-In Utilities – A set of utilities used by the other NAG Books
 Book 0 – Support functions for elementary statistical calculation
 Book 1 – Statistical Modelling 1: Regression and Time Series
 Book 2 – Statistical Modelling 2: Analysis of Experiments and Generalized Linear Models
 Book 3 – Multivariate Methods
 NAGExtract – A utility to reposition output more conveniently on the worksheet

5.2. *Book 0*

5.2.1. Introduction

Introduction to the Basic Statistics

Book 0 provides a range of basic statistics and probability functions. These are generally intended to support the functions in the other books. While much of the functionality is already available in Excel there are advantages with using a common source for your statistical analysis. The functions can be split into three categories:

1. Distribution functions
2. Summary statistics
3. Contingency tables

Please note that a complete set of examples for Book 0 are provided in the workbook Book0E.xls. This is installed in the 'Examples' subdirectory of the Add-Ins directory. You might find it helpful to work through these examples to familiarize yourself with the use of the Book 0 Add-Ins.

5.2.2. Distribution Functions

5.2.2.1. Introduction

Distribution functions

Functions that calculate the probability for a value and functions that calculate the value from the probability (percentage points/deviates) are available for the following distributions:

Distribution	Probability	Deviate
Normal	NORMAL_PROB	NORMAL_PPT
t	T_PROB	T_PPT
χ^2	CHI_PROB	CHI_PPT
F	F_PROB	F_PPT
Gamma	GAMMA_PROB	GAMMA_PPT
Beta	BETA_PROB	BETA_PPT

Notes: The probability density function of the gamma distribution is given by

$$\exp((a - 1)\log(x/b) - x/b)/(b\Gamma(a))$$

where a is the shape parameter and b is the scale parameter, with the mean given by ab .

The beta distribution is also known as the beta distribution of the first kind.

The degrees of freedom for the t , F and χ^2 distributions may be non-integer.

The accuracy of these probability functions varies with their intended use. For GAMMA_PPT, BETA_PROB and BETA_PPT, there is an option to set the required accuracy. For NORMAL_PROB, NORMAL_PPT and GAMMA_PROB, the accuracy should be about 16 decimal places. For the remainder, the accuracy is about 5 decimal places for normal usage. Further information can be found in the NAG Fortran 77 Library Documentation.

In addition to the above univariate distributions there is also a function that returns the probability for the multivariate normal distribution. This requires the input of the variance-covariance matrix for the distribution as well as its mean vector.

Several of the NAG distribution functions have an option for the selection of the tail of the distribution. The values available will depend on the function but the possibilities are

1. Lower tail
2. Upper tail
3. Two tail for significance testing
4. Two tail for confidence intervals

The two tail options are only available for the symmetric distributions Normal and t -distribution. For the significance testing option the sum of the two equal tail probabilities is used; that is 0.05 would be used for a two tailed test at the 5% level. In the case of the confidence interval option the central symmetric probability is used; that is 0.95 would be used for a 95% confidence interval.

Unlike the rest of the NAG functions, the distribution functions return a single value rather than an array of values, so the NAGExtract utility is not required.

5.2.2.2. NORMAL_PROB (Set 1)

NORMAL_PROB (NAG G01EAF) returns, as a single value, a one or two tail probability for the standard Normal distribution.

Parameters

X [Required]
The value of the standard Normal variate.

Tail [Required]

Indicates which tail the return probability should represent.

- U: The upper tail probability is returned.
- L: The lower tail probability is returned.
- S: The two tail (significance level) is returned.
- C: The two tail (confidence interval) probability is returned.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.3. NORMAL_PPT (Set 1)

NORMAL_PPT (NAG G01FAF) returns, as a single value, the deviate associated with the given probability of the standard Normal distribution.

Parameters

Probability [Required]

The probability of the standard Normal distribution as defined by tail.

Tail [Required]

Indicates which tail the supplied probability represents.

- U: The upper tail probability.
- L: The lower tail probability.
- S: The two tail (significance level) probability.
- C: The two tail (confidence interval) probability.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.4. MULTI_NORMAL_PROB (Set 1)

MULTI_NORMAL_PROB (NAG G01HBF) returns, as a single value, the upper tail, lower tail or central probability associated with a multivariate Normal distribution of up to ten dimensions.

Parameters

Mean [Required]

The mean vector of the multivariate Normal distribution.

Covariance [Required]

The square array containing the variance-covariance matrix of the multivariate Normal distribution. Only the lower triangle is referenced.

Lower_bounds [Optional]

If Tail = 'C' or 'U', the lower bounds.

Upper_bounds [Optional]

If Tail = 'C' or 'L', the upper bounds.

Tail [Optional]

Indicates which probability is to be returned.

- U: The upper tail probability.
 - L: The lower tail probability.
 - C: The central probability is returned.
- Default: central probability

Tolerance [Optional]

The relative accuracy required in the result.

Default: 0.0001

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.5. T_PROB (Set 1)

T_PROB (NAG G01EBF) returns, as a single value, the lower tail, upper tail or two tail probability for the Student's t -distribution.

Parameters

T [Required]

The value of the Student's t variate.

Df [Required]

The degrees of freedom of the Student's t -distribution, must be ≥ 1 .

Tail [Required]

Indicates which tail the return probability should represent.

- U: The upper tail probability is returned.
- L: The lower tail probability is returned.
- S: The two tail (significance level) is returned.
- C: The two tail (confidence interval) is returned.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.6. T_PPT (Set 1)

T_PPT (NAG G01FBF) returns, as a single value, the deviate associated with the given tail probability of Student's t -distribution.

Parameters

Probability [Required]

The probability of the required Student's t -distribution as defined by tail.

Df [Required]

The degrees of freedom of the Student's t -distribution, must be ≥ 1.0 .

Tail [Required]

Indicates which tail the supplied probability represents.

- U: The upper tail probability.
- L: The lower tail probability.
- S: The two tail (significance level) probability.
- C: The two tail (confidence interval) probability.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.7. CHI_PROB (Set 1)

CHI_PROB (NAG G01ECF) returns, as a single value, the lower or upper tail probability for the χ^2 distribution.

Parameters

X [Required]

The value of the χ^2 variate, must be > 0 .

Df [Required]

The degrees of freedom of the χ^2 distribution, must be > 0 .

Tail [Required]

Indicates which tail the return probability should represent.

- U: The upper tail probability is returned.
- L: The lower tail probability is returned.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.8. CHI_PPT (Set 1)

CHI_PPT (NAG G01FCF) returns, as a single value, the deviate associated with the given lower tail probability of the χ^2 distribution.

Parameters**Probability [Required]**

The probability of the required χ^2 distribution.

Df [Required]

The degrees of freedom of the χ^2 distribution, must be > 0.0 .

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.9. F_PROB (Set 1)

F_PROB (NAG G01EDF) returns, as a single value, the lower or upper tail of the F or variance-ratio distribution.

Parameters**X [Required]**

The value of the F variate, must be ≥ 0.0 .

Numerator_df [Required]

The degrees of freedom of the numerator variance, must be > 0.0 .

Denominator_df [Required]

The degrees of freedom of the denominator variance, must be > 0.0 .

Tail [Required]

Indicates whether an upper or lower tail probability is required.

- U: The upper tail probability is returned.
- L: The lower tail probability is returned.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.10. F_PPT (Set 1)

F_PPT (NAG G01FDF) returns, as a single value, the deviate associated with the given lower tail probability, of the F or variance-ratio distribution.

Parameters**Probability [Required]**

The probability from the required F-distribution.

Numerator_df [Required]

The degrees of freedom of the numerator variance, must be > 0.0 .

Denominator_df [Required]

The degrees of freedom of the denominator variance, > 0.0 .

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.11. BETA_PROB (Set 1)

BETA_PROB (NAG G01EEF) returns, as a single value, the lower or upper tail probability of the beta distribution.

Parameters**X** [Required]

The value of the beta variate, $0.0 \leq X \leq 1.0$.

A [Required]

The first parameter of the required beta distribution, $0.0 < A \leq 10^6$.

B [Required]

The second parameter of the required beta distribution, $0.0 < B \leq 10^6$.

Tail [Required]

Indicates whether an upper or lower tail probability is required.

U: The upper tail probability is returned.

L: The lower tail probability is returned.

Tolerance [Optional]

The relative accuracy required by the user in the results.

Default: a value of 10 times the machine precision is used.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.12. BETA_PPT (Set 1)

BETA_PPT (NAG G01FEF) returns, as a single value, the deviate associated with the given lower tail probability of the beta distribution.

Parameters**Probability** [Required]

The probability from the required beta distribution.

A [Required]

The first parameter of the required beta distribution, $0.0 < A \leq 10^6$.

B [Required]

The second parameter of the required beta distribution, $0.0 < B \leq 10^6$.

Tolerance [Optional]

The relative accuracy required in the result.

Default: 10 times machine precision.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.13. GAMMA_PROB (Set 1)

GAMMA_PROB (NAG G01EFF) returns, as a single value, the lower or upper tail probability of the gamma distribution.

Parameters

X [Required]

The value of the gamma variate, must be ≥ 0.0 .

Shape_param [Required]

The shape parameter of the gamma distribution, must be > 0.0 .

Scale_param [Required]

The scale parameter of the gamma distribution must be > 0.0 .

Tail [Required]

Indicates which tail the return probability should represent.

U: The upper tail probability is returned.

L: The lower tail probability is returned.

Remarks

The parameterization of the gamma distribution used by this function is such that if the shape parameter is a and the scale parameter is b the mean is ab and the probability density function is

$$\exp((a - 1)\log(x/b) - x/b)/(b\Gamma(a)).$$

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.2.14. GAMMA_PPT (Set 1)

GAMMA_PPT (NAG G01FFF) returns, as a single value, the deviate associated with the given lower tail probability of the gamma distribution

Parameters

Probability [Required]

The probability from the required gamma distribution.

Shape_param [Required]

The shape parameter, a , of the gamma distribution, $a > 0.0$.

Scale_param [Required]

The scale parameter, b , of the gamma distribution, $b > 0.0$.

Tolerance [Optional]

The relative accuracy required in the result.

Default: 10 times machine precision.

Remarks

The parameterization of the gamma distribution used by this function is such that if the shape parameter is a and the scale parameter is b the mean is ab and the probability density function is

$$\exp((a - 1)\log(x/b) - x/b)/(b\Gamma(a))$$

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.3. Summary Statistics

5.2.3.1. Introduction

Summary Statistics

Two functions are available for the calculation of basic summary statistics. SUMMARY_STATS provides the mean, standard deviation, and coefficients of skewness and kurtosis, as well as the maximum and minimum values for a set of ungrouped data. The function S5PT_SUMMARY provides the median along with the maximum, minimum and the upper and lower hinges. The hinges are virtually the same as the quartiles. The five point summary is the basis of a box and whisker plot but can also be used directly to assess transformations for symmetry.

Contingency Tables

A two-way contingency table arises when a sample has been classified by two qualitative variables, for example, sex (male / female) and occupation type (managerial / professional / skilled / unskilled). The numbers in each of the resulting 2 by 4 categories form the cells of the two-way contingency table (these observed numbers are denoted by O_{ij}). The standard analysis is the chi-squared test. This computes what are the expected numbers (denoted by E_{ij}) in each cell if there was no association between the two qualitative variables that define the table. The resulting statistic is

$$X^2 = \sum [(O_{ij} - E_{ij})^2 / E_{ij}]$$

where the sum is over all cells. An alternative statistic is the likelihood ratio test statistic given by

$$G^2 = 2 \sum O_{ij} \log(O_{ij}/E_{ij}).$$

In the case when both qualitative variables have just two levels, the chi-squared statistic for the resulting 2×2 table can be improved by Yates' correction factor and for 2×2 tables with small numbers in each cell Fisher's exact test can be computed. The function TWO_WAY_TABLE computes both the chi-squared and likelihood ratio statistics for a two-way table. If the total sample size in a 2×2 table is less than or equal to 40 Fisher's exact test is performed. If the expected value of a cell is less than 0.5 a warning is given since the distribution of the resulting chi-squared statistic may not be well approximated by the chi-squared distribution.

For further information see, for example, Everitt (1977).

5.2.3.2. SUMMARY_STATS (Set 1)

SUMMARY_STATS (NAG G01AAF) calculates the mean, standard deviation, coefficients of skewness and kurtosis, and the maximum and minimum values for a set of ungrouped data. Weighting may be used.

Parameters**X** [Required]

The observations upon which the calculations are to be performed.

Weights [Optional]

If present, Weights must contain the non-negative weights associated with the observations.

OutputOutput as a **NAGExtract format list**.**Number of valid cases****Mean****Standard Deviation****Coefficient of Skewness****Coefficient of Kurtosis****Minimum****Maximum**

If weights are used then also

Sum of Weights

Note:

When weights are used, the number of valid cases is the number of observations with non-zero weights, otherwise it is simply the number of observations.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.3.3. S5PT_SUMMARY (Set 1)

S5PT_SUMMARY (NAG G01ALF) calculates a five-point summary for a single sample.

Parameters**X** [Required]

The sample; there must be at least 5 observations.

OutputOutput as a **NAGExtract format list**.**Minimum****Lower Hinge****Median****Upper Hinge****Maximum****Remarks**

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.2.3.4. TWO_WAY_TABLE (Set 1)

TWO_WAY_TABLE (NAG G11AAF) computes chi-squared statistics for a two-way contingency table. For a 2×2 table with 40 or fewer observations an exact probability is computed.

Parameters**Table** [Required]

The array containing the contingency table. Margins should not be included.

Output**Significance****Chi squared****Likelihood ratio****Degrees of freedom****Expected values****Chi-squared contributions****Notes:**

The significance returned is for the chi-squared statistic except for a 2×2 table when the total number of observations is less than or equal to 40, in which case the probability from Fisher's exact test is returned.

The chi-squared contributions are the values of $(\text{observed} - \text{expected})^2 / \text{expected}^2$ for each cell which make up the chi-squared statistic.

Remarks

See 'Using the Add-Ins' and the Introduction to Basic Statistics for further information.

5.3. Book 1

5.3.1. Introduction

This book contains materials for two areas:

- Introduction to Regression and Correlation
- Introduction to Time-Series

Please note that a complete set of examples for this Add-In are provided in the workbook Book1E.xls. This is installed in the 'Examples' subdirectory of the Add-Ins directory. You might find it helpful to work through this to familiarize yourself with the use of the Book 1 Add-Ins.

5.3.2. Regression and Correlation

5.3.2.1. Introduction

Introduction to Regression and Correlation

Regression and correlation are both concerned with the relationship between variables. In the case of correlation, the variables are considered in pairs and the correlation coefficient (sometimes known as the Pearson product-moment correlation coefficient) measures the linear association between the two variables, that is, how close the plotted points would be to lying on a straight line. The correlation coefficient lies between -1 and $+1$, with 0 indicating no linear association between the two variables and ± 1 indicating a perfect linear relationship so that the plotted points would lie on a straight line with positive slope (correlation of 1) or negative slope (correlation of -1). The function CORREL_MAT computes all the pairs of correlations for a set of variables, known as a correlation matrix. Correlation can also be seen as a standardized form of the covariance between two variables. CORREL_MAT also returns the variance-covariance matrix.

Sometimes the correlation between two variables should not be interpreted as a direct relationship between the two variables but as being due, in part at least, to their being correlated with a third variable. For example, using historical data, there may be a significant correlation between electricity consumption and road accidents. This would mainly be due to both increasing over time. To remove the effect of the third (or other) variables, a partial correlation coefficient can be computed. This measures the linear association between two variables having eliminated the linear effect of other variables. The function PARTIAL_CORREL_MAT computes partial correlations.

In contrast to correlation, regression looks at how one variable is linearly related to other variables. This variable is known as the dependent or response variable and is often denoted by y . The other variables are known as the independent or predictor variables and are often denoted by x_i . In the case of one independent variable the model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where ε is an independent error or noise term with mean zero and constant variance σ^2 . This can be extended to many variables, x_1, x_2, x_3 etc., for example,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

which is known as a multiple linear regression model. A particular case of this is the polynomial regression model in which $x_2 = x_1$ squared, etc., giving

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon.$$

In cases where, instead of continuous variables, there are classification or grouping variables, the multiple linear regression model is known as the general linear model, see Searle (1971). In the general linear model the classification variables are fitted by means of a number of dummy (0,1) variables which represent each class or level. For example, if the classification variable is colour and takes one of the values: red, green, blue or brown, the following dummy variables can be used.

Classification Variable	Dummy Variable 1	Dummy Variable 2	Dummy Variable 3	Dummy Variable 4
Red	1	0	0	0
Green	0	1	0	0
Green	0	1	0	0
Blue	0	0	1	0
Brown	0	0	0	1
Red	1	0	0	0
Blue	0	1	0	0

As the sum of the dummy variables is one for each observation, the set of dummy variables will not be of full rank if a mean term is fitted (equivalent to 1 for each observation). So usually dummy variable 1 is not included so that the other dummy variables represent comparisons with this first class, in this case red. Alternatively, the comparison can be with the last or the average of the previous classes (Helmert matrix). When each level represents some underlying quantitative variable an orthogonal polynomial representation can be used.

In addition to the effect of a single classification variable there can be the effect of the combination of two classification variables. For example, if there were also a classification by sex then there would be an effect due to colour, and an effect due to sex plus an effect due to

the combination of a sex and colour. For instance, if females were in general larger than males and reds were, in general, larger than blues, greens and browns it may be, however, that green males were even smaller than expected or that blue males were bigger than red females. Such effects are known as interactions and indicate that the combination of colour and sex should be examined rather than just examining one variable at a time.

The function `MULT_LIN_REG` fits a multiple linear regression model. There are two ways of specifying the independent variables. One way is to provide an array with the variables stored in columns. This information can then be supplemented by optional names for the variables and 0/1 indicators that allow variables to be dropped from the model. Dummy variables can be calculated using `DUMMY_VARS` but a more general way of specifying the model is to use the `MM_MLR` function described below. This uses an algebraic specification of the model that is found in most statistical packages and automatically generates the required dummy variables.

The regression fitting returns an analysis of variance table, parameter estimates with their standard errors, *t*-values and correlations along with fitted values, residuals and leverages. The leverages provide an indication of how much influence the observation has on the fit and they can also be used to compute confidence intervals for the fitted line. Further examination of the residuals can be assisted by the use of standardized residuals computed by `STAND_RESIDS`, which also computes the Cook's D statistics that measure the influence of an observation; see Cook and Weiberg (1982). If the observations have been observed serially, `DURBIN_WATSON` computes the Durbin–Watson test for serial correlation.

Specifying the Model Matrix

The NAG function `MM_MLR` can be used to specify a model matrix rather than the user constructing a set of columns. The inputs to the model matrix function `MM_MLR` are

1. Group variables with their names
2. Covariates with their names
3. A formula

where, as described above, the group variables are classification variables, e.g., eye colour, sex. These variables can be either text, e.g., red, green, blue, or numerical, e.g., 100, 150, 300. `MM_MLR` recodes these values first into an integer representation, 1,2,..., then into the dummy variables required for the model fitting. Covariates are ordinary variables, e.g., weight, temperature.

A formula uses the variable names and the symbols

$+ - * . / () ^$

Some examples of formulae are:

Group variables: a, b and c

Covariates: x and z

$a + x$ fits the group variable a and covariate x, that is, regression model with different intercepts for the different groups.

$x + x^2$ fits the quadratic model in x.

$a * b$ fits the terms $a + b + a.b$, that is, the main effects of a and b and the a.b interaction, that is a cross-classification model.

a / b fits $a + a.b$, that is the nested model.

$(a+b)*c$ fits $a*c$ plus $b*c$, that is, $a + b + c + a.c + b.c$

The ^ notation is only available for use with covariates. There is also an additional parameter, Interaction, which sets the maximum number of terms allowed in an interaction. So, for example,

$a*b*c$ with Interaction = 2 gives $a + b + c + a.b + a.c + b.c$ with the three-term interaction excluded.

Note that the model matrix produced by MM_MLR assumes that the mean is also fitted. If the mean is to be omitted then DUMMY_VARS should be used to construct the required dummy variables with all levels present.

5.3.2.2. CORREL_MAT (Set 1)

CORREL_MAT (NAG G02BXF) calculates the product-moment correlation matrix and the variance-covariance matrix.

Parameters

X [Required]

The array containing the observations on the variables in columns.

One of the following may be set:

Frequencies [Optional]

The number of observations associated with each observation. The effective number of observations is taken as the sum of the frequencies.

Weights [Optional]

The weights associated with each observation. The number of observations is taken as the number of observations with non-zero weights.

In either case the value zero means that an observation is excluded from the analysis.

Output

Output as a NAGExtract format list.

Variance-Covariance Matrix

Correlation Matrix

Mean

Standard Deviation

Remarks

See 'Using the Add-Ins' and the Introduction to Regression and Correlation for further information.

5.3.2.3. PARTIAL_CORREL_MAT (Set 1)

PARTIAL_CORREL_MAT (NAG G02BYF) computes a partial correlation/variance-covariance matrix from a correlation or variance-covariance matrix computed by CORREL_MAT.

Parameters

Correlations [Required]

Correlation matrix (or the variance-covariance matrix).

Select_vars [Required]

Indicates which variables belong to set X and Y.

- Y: Y or dependent variable.
- X: X or independent variable.
- O: variable is omitted.

Output

Output as a NAGExtract format list.

Partial Correlation Matrix
Partial Covariance Matrix

Note:

Both are returned as upper triangular matrices. The partial covariance matrix contains the true partial variance-covariance matrix if a variance-covariance matrix is supplied and a variance-covariance matrix for standardized variables if a correlation matrix is supplied.

Remarks

See 'Using the Add-Ins' and the Introduction to Regression and Correlation for further information.

5.3.2.4. MULT_LIN_REG (Set 1)

MULT_LIN_REG (NAG G02DAF) fits a general multiple linear regression model.

Parameters**Y** [Required]

Observations on the dependent variable.

X [Required]

Either the array containing the independent variables as columns or the function MM_MLR that computes the model matrix.

Note that X_names and Select_x should not be used if MM_MLR is used.

Select_x [Optional]

Indicates which independent variables in array X are to be included in the model. If $\text{Select_x}(j) > 0$, then the variable contained in the j th column of array X is included in the regression model, if $\text{Select_x}(j) = 0$ it is excluded. If present, Select_x values ≥ 0 .

Default: all X variables are included.

X_names [Optional]

Names for the independent variables in array X.

Default: X_1, X_2, etc.

Mean [Optional]

Indicates whether a mean term is to be included.

- M: A mean term, intercept, will be included in the model.
- Z: The model will pass through the origin, zero point.

Weights [Optional]

Contains the optional weights to be used in the regression. A zero weight indicates that an observation is excluded from the analysis. If present, $\text{Weights} \geq 0.0$.

Fitted_values [Optional]

Indicates whether fitted values, residuals and leverages are to be returned.

Y: Fitted values etc. are returned.

N: Fitted values etc. are not returned.

Default: Y – Fitted values etc. are returned.

Tolerance [Optional]

The value of Tolerance is used to decide if the independent variables are of full rank, and if not, what is the rank of the independent variables. If present, $\text{Tolerance} > 0.0$.

Default: 0.000001.

Output

Output as a NAGExtract format list. Some or all of the following may be returned.

Rank of Model**ANOVA****R-squared****Parameter Estimates****Fitted Values****Residuals****Leverages****Covariance Matrix****Notes:**

The rank of the model is only returned if it is not of full rank.

The ANOVA table for the model is only returned if the residual degrees of freedom are not zero, as is the Covariance Matrix and the standard errors etc. for the parameters.

The Parameter Estimates array contains

Parameter Names

Estimates

Standard Errors

T-values

Significance

The Fitted Values, Residuals and Leverages are not returned if $\text{Fitted_values} = \text{N}$.

Remarks

See 'Using the Add-Ins' and the Introduction to Regression and Correlation for further information.

5.3.2.5. MM_MLR (Set 1)

MM_MLR computes the model matrix for a general linear model specified by a formula. It is for use with MULT_LIN_REG, where it is called instead of providing the X array.

Parameters**Formula** [Required]

The model formula that may consists of the names of the variables given in Group_Names and Covariate_Names and the following symbols:

() + . * / - ^

where the symbols represent the following actions

- + add term to formula
- . interaction between two variables
- * cross-classify two terms, e.g., $a*b = a + b + a.b$
- / nest two terms, e.g. $a/b = a + a.b$
- remove term from formula, e.g., $a*b - a.b = a + b$
- ^ power for covariates only

The precedence of the operators is () - + * / . ^

At least one of Groups or Covariates must be given along with the corresponding names.

Groups [Optional]

The array whose columns contain the variables that are to be considered as grouping or classification variables.

Group_Names [Optional]

The names of the variables in Groups.

Covariates [Optional]

The array whose columns contain the variables that are to be considered as covariates, i.e., not as classification variables.

Covariate_Names [Optional]

The names of the variables in Covariates.

Interaction [Optional]

The maximum number of terms to be included in an interaction, ≥ 1 .

Default: all terms included.

Output

To avoid the restriction on the size of returned arrays the function MM_MLR passes the results to the calling function by means of global variables. All that is returned directly is an indicator string or any error messages.

The function computes all required dummy variables with appropriate names. The dummy variable corresponding to the first level of a group variable is not included in order to provide a full rank model.

Remarks

See 'Using the Add-Ins' and the Introduction to Regression and Correlation for further information.

5.3.2.6. DURBIN_WATSON (Set 1)

DURBIN_WATSON (NAG G02FCF) calculates the Durbin–Watson test for serial correlation in linear regression.

Parameters

No_of_params [Required]

The number of independent variables in the regression model, including the mean.

Residuals [Required]

The residuals from the regression. Note that the mean of the residuals must be close to zero (approximately $\pm 10^{-8}$).

Output

Output as a NAGExtract format list.

Durbin Watson Statistic

Lower Bound for Significance

Upper Bound for Significance

Note:

The Durbin–Watson test does not have an exact level of significance, only an upper or lower bound. The statistics returned are for the test for positive auto-correlation; for negative auto-correlation the value 4 – the Durbin–Watson statistic should be calculated and compared with tables.

Remarks

See ‘Using the Add-Ins’ and the Introduction to Regression and Correlation for further information.

5.3.2.7. STAND_RESIDS (Set 1)

STAND_RESIDS (NAG G02FAF) calculates two types of standardized residuals and two measures of influence for a linear regression.

Parameters

Residuals [Required]

The residuals from the regression.

Leverages [Required]

The leverages from the regression.

RMS [Required]

The residual mean square from the regression.

RDF [Required]

The residual degrees of freedom from the regression.

Output

Output as a NAGExtract format list.

Internally Standardized Residuals

Externally Standardized Residuals

Cook’s D

Atkinson’s T

Remarks

See 'Using the Add-Ins' and the Introduction to Regression and Correlation for further information.

5.3.2.8. DUMMY_VARS (Set 1)

DUMMY_VARS (NAG G04EAF) computes orthogonal polynomial or dummy variables for a factor or classification variable.

Parameters**Factor** [Required]

The integer values of the factor, which must be between 1 and Levels.

Levels [Required]

The number of levels (number of classes) of the factor. There must be at least two levels.

Form [Optional]

The form of the dummy variable to be computed.

- P: an orthogonal Polynomial representation is computed.
- H: a Helmert matrix representation is computed.
- F: the contrasts relative to the first level are computed.
- L: the contrasts relative to the last level are computed.
- C: a complete set of dummy variables is computed.

Default: contrasts relative to the first level.

X_values [Optional]

If Form = 'P', the distinct values, corresponding to the levels, of the underlying variable for which the orthogonal polynomial is to be computed.

Output

Output as a NAGExtract format list.

X

Note:

The array of dummy variables returned, X, will have Levels – 1 columns if Form is set to P, H, F or L and Levels columns if it is set to C.

Remarks

DUMMY_VARS should be used when an orthogonal polynomial or a specific form for the dummy variables is required. For general use the NAG MM_MLR function computes the dummy variables required for a regression model from the formula provided.

See 'Using the Add-Ins' and the Introduction to Regression and Correlation for further information.

5.3.3. Time-Series

5.3.3.1. Introduction

Introduction to Time Series Methods

Time Series Models

The time series considered in this book are sequences of observations observed at equally spaced intervals, for example, the temperature of a process recorded every minute, the daily price of shares or the monthly customer numbers. For an introduction to time series see Chatfield (1982). The time series can be considered as having three basic components:

1. Trend
2. Cyclic or periodic
3. Random or noise

The trend represents the long term pattern, e.g., increasing sales, while the cyclic or periodic components follow patterns such as 'summer sales are higher than winter sales' or 'there are more enquires on a Monday morning than on a Friday afternoon'. The random component is the fluctuation around the trend and cycles. Unlike regression models, the noise in time series models is not assumed to be independent but is (auto) correlated. Furthermore, the noise is usually assumed to have zero mean and to be stationary. Being stationary means that the statistical relationship between the observation at time t and time $t+l$ is the same as the statistical relationship between observations at time $t+t$ and $t+l+\tau$. The relationship can be described in terms of either their joint distribution or in terms of correlation. The theoretical correlation between an observation at time t and an observation at time $t+l$, that is at lag l , is denoted by ρ_l .

Statistical models for time series relate the correlated noise (denoted by w_t) to an independent random component or white noise (denoted by ε_t), and hence explain the autocorrelation r_l . The most common models for stationary time series are autoregressive models and moving average models. An autoregressive model for a series w_t of order p is

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \dots + \phi_p w_{t-p} + \varepsilon_t$$

which is denoted by AR(p). A moving average model of order q is

$$w_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}$$

which is denoted by MA(q). The difference between the two forms of model is that the autocorrelation for an autoregressive model gradually dies away while the autocorrelation for a moving average model will cut off at lag q .

In the situation where the data has a cyclic or periodic nature, for example days of the week, a seasonal model may be appropriate. For example, Tuesday's observation may be related to both Monday's observation and the previous Tuesday's observation. In seasonal models the terms are related to the seasonality, s , e.g., for days of the week $s = 7$. The seasonal autoregressive and moving average models are

$$w_t = \Phi_1 w_{t-s} + \Phi_2 w_{t-2s} + \dots + \Phi_P w_{t-Ps} + \varepsilon_t$$

and

$$w_t = \varepsilon_t - \theta_1 \varepsilon_{t-s} - \theta_2 \varepsilon_{t-2s} - \dots - \theta_Q \varepsilon_{t-Qs}.$$

These models can be combined to give a seasonal ARMA model

$$w_t = \Phi_1 w_{t-s} + \Phi_2 w_{t-2s} + \dots + \Phi_P w_{t-Ps} + e_t - \theta_1 e_{t-s} - \theta_2 e_{t-2s} - \dots - \theta_Q e_{t-Qs}$$

$$e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \dots + \phi_p e_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}$$

where e_t is an intermediate series for the purposes of defining the model.

In the situation where the series y_t is not stationary it may be possible to make it stationary by applying differencing or seasonal differencing. First-order differencing is given by

$$\nabla y_t = y_t - y_{t-1}$$

and first-order seasonal differencing is given by

$$\nabla_s y_t = y_t - y_{t-1}.$$

Differencing can be combined with a constant term to give

$$\nabla^d \nabla_s^D y_t = c + w_t.$$

Combining this with the ARMA model given above gives a seasonal ARIMA (AutoRegressive Integrated Moving Average) model which can be specified by the seven values

$$(p, d, q, P, D, Q, s)$$

Time Series Model Identification

The basic tools in the identification of a suitable time series model are data plots, the autocorrelation function (acf) and the partial autocorrelation function (pacf). Plotting the data will indicate if the series is stationary. If not, the series can be transformed using standard functions such as LOG or SQRT and differencing can be applied using the function TIME_SERIES_DIFF.

Given a stationary time series, the acf and pacf can be calculated using the functions ACF and PACF respectively. By looking at the acf the order, q , of a possible MA model can be identified, while the pacf for an AR(p) model will cut off at lag p . The function PACF also gives approximate estimates of the parameters of the autoregressive model of order given by the number of partial autocorrelations requested and the predictor error variance ratio which is

$$V_I = \text{var}(\varepsilon_{I,t}) / \text{var}(w_t)$$

where $\text{var}(w_t)$ is the variance of the stationary series and $\text{var}(\varepsilon_{I,t})$ is the estimated variance of the white noise when an AR(I) has been fitted.

If neither approach is suitable then ARMA models can be considered. The function ARIMA_APPROX_FIT gives an approximate fit for an ARIMA model. This may be useful in model identification or for giving a set of initial values for the full fitting process.

Time Series Model Fitting and Forecasting

The function ARIMA_FIT fits an ARIMA model using either maximum likelihood or least squares. The model is specified by the orders p , d , q , P , D , Q and s . The function returns the parameter estimates, standard errors, t -values and significance. The results can be input into ARIMA_FORECAST to forecast values for the series.

Transfer Function Models

The ARIMA model can be used to describe a single series; however, if the (output) series can be related to several input series then a multi-input or transfer function model can be used. The form of a transfer function model is

$$y_t = z_{1,t} + z_{2,t} + \dots + z_{m,t} + n_t$$

where n_t follows an ARIMA model and the components of the model, $z_{i,t}$, are related to the input variables, $x_{i,t}$, by either a simple linear model,

$$z_{i,t} = \omega_0 x_{i,t}$$

or an ARMA-like model

$$z_{i,t} = \delta_1 z_{i,t-1} + \delta_2 z_{i,t-2} + \dots + \delta_p z_{i,t-p} + \omega_0 x_{i,t-b} + \omega_1 x_{i,t-b-1} + \dots + \omega_q x_{i,t-b-q}$$

where b is known as the delay. A transfer function model can be fitted using the options of ARIMA_FIT, and forecasts can be made using TRNS_FUNC_FORECAST. To forecast for the output series of a transfer function model the forecasts of the input series have to be provided. Also, if the input series have been forecast using ARIMA models, these models may be supplied to TRNS_FUNC_FORECAST; this will not affect the forecast but will adjust the forecast standard error.

Spectral Analysis

An alternative way of looking at a time series is to consider it as being made up of a number of sine/cosine waves of different frequency. The (power) spectrum indicates how much of the variance of the series can be attributed to the different frequencies. A plot of the spectrum against the frequency is known as a periodogram. Spectral analysis is useful in identifying cyclic components in a series.

The computed sample spectrum can be smoothed to aid interpretation or to provide an estimate of the underlying spectral density. The function SPECTRAL smoothes the sample spectrum it calculates by using a trapezium frequency window, known as a Daniell Window. The shape of the window can vary from triangular to rectangular and the amount of smoothing is controlled by the width of the window. While the unsmoothed sample spectrum ordinates are independent for the smoothed spectrum they are only independent if they are the amount given by the bandwidth statistic apart. The distribution of the spectrum is approximately that of a scaled χ^2 variate with degrees of freedom as returned by SPECTRAL. This approximate distribution gives a multiplicative (or additive if the log-spectrum is selected) confidence interval for the spectrum.

The function SPECTRAL computes the sample spectrum by means of a fast Fourier transform (FFT). For efficiency the length of the series passed to the FFT should be at least twice the length of the observed series and a product of small prime numbers. To achieve this with a series of arbitrary length the series is extended with zeros until it is of the required length. In order to provide a smooth transition to the zero values the data can be tapered at both ends.

The spectrum is then calculated for a subset of the values computed by the FFT as given by the frequency division. A divisor of 2π is used when computing the sample spectrum from the FFT.

5.3.3.2. TIME_SERIES_DIFF (Set 1)

TIME_SERIES_DIFF (NAG G13AAF) carries out non-seasonal and seasonal differencing on a time series.

Parameters

Series [Required]

The time series.

Note:

At least one of Differencing or Seasonal_diff must be present.

Differencing [Optional]

The order of non-seasonal differencing. If present, must be > 0 .

Seasonal_diff [Optional]

The order of seasonal differencing. If present, must be > 0 .

Seasonality [Optional]

The seasonality for the seasonal differencing. Must be present if Seasonal_diff is present. If present, must be > 1 .

Output

Output as a NAGExtract format list.

Differenced Series

Remarks

See 'Using the Add-Ins' and the Introduction to Time Series Methods for further information.

5.3.3.3. ACF (Set 1)

ACF (NAG G13ABF) computes the sample autocorrelation function of a time series along with a test for autocorrelation.

Parameters

Series [Required]

The time series.

No_of_autocorrels [Optional]

The number of lags for which the autocorrelations are required.

Default: $\min(20, \text{number of observations} - 1)$.

Output

Output as a NAGExtract format list.

Mean

Variance

Chi squared Statistic

Chi squared Significance

ACF

Remarks

See 'Using the Add-Ins' and the Introduction to Time Series Methods for further information.

5.3.3.4. PACF (Set 1)

PACF (NAG G13ACF) calculates the partial autocorrelation function from the autocorrelation function.

Parameters

ACF [Required]

The autocorrelation function.

Output

Output as a NAGExtract format list.

PACF

Predictor Error Variance Ratio

Autoregressive Parameters

Note:

The number of partial autocorrelations computed is the maximum that can be computed from the supplied ACF; this may be less than the number of autocorrelations. The Predictor Error Variance Ratios are defined in the Introduction to Time Series Methods. The Autoregressive parameters are for the maximal autoregressive model that can be computed from the autocorrelations.

Remarks

See 'Using the Add-Ins' and the Introduction to Time Series Methods for further information.

5.3.3.5. ARIMA_APPROX_FIT (Set 1)

ARIMA_APPROX_FIT (NAG G13ADF) calculates preliminary estimates of the parameters of an autoregressive integrated moving average (ARIMA) model from the autocorrelation function of the appropriately differenced time series.

Parameters

Model [Required]

Model specifies the orders (p, d, q, P, D, Q, s) of the (seasonal) ARIMA model to be fitted.

- p: number of autoregressive terms.
- d: order of non-seasonal differencing.
- q: number of moving average terms.
- P: number of seasonal autoregressive terms.
- D: order of seasonal differencing.
- Q: number of seasonal moving average terms.
- s: the seasonality.

Variance [Required]

The series sample variance, calculated after appropriate differencing has been applied to the series. See output of ACF.

ACF [Required]

The autocorrelations (starting at lag 1), which must have been calculated after the time series has been appropriately differenced. See output of ACF.

Output

Output as a NAGExtract format list.

Residual Variance Parameters

Note:

The parameters are returned in the order autoregressive, moving average, seasonal autoregressive and seasonal moving average.

Remarks

See 'Using the Add-Ins' and the Introduction to Time Series Methods for further information.

5.3.3.6. ARIMA_FIT (Set 1)

ARIMA_FIT (NAG G13BEF) fits either a (seasonal) ARIMA model or a multi-input (transfer function) model.

Parameters

Y_series [Required]

The (output) time series.

Model [Required]

Model specifies the orders (p, d, q, P, D, Q, s) of the (seasonal) ARIMA model to be fitted to the output series.

- p: number of autoregressive terms.
- d: order of non-seasonal differencing.
- q: number of moving average terms.
- P: number of seasonal autoregressive terms.
- D: order of seasonal differencing.
- Q: number of seasonal moving average terms.
- s: the seasonality.

Fix_constant [Optional]

Indicates if the constant is to be fixed at its initial value or estimated.

- F: fixed at initial values.
- E: estimated.

Default: E – estimated.

Parameters [Optional]

Initial values of the model parameters in the order: ARMA parameters (autoregressive, moving average, seasonal autoregressive and seasonal moving average), transfer function parameters (omega parameters followed by delta parameters), if present, and finally the constant.

Default: initial parameters are set to zero except for the constant, which is set equal to the output series mean

T_Model [Optional]

For a multi-input (transfer function) model the *j*th column of the array T_Model specifies the form of the model relating the *j*th input series to the output series. The array must have 4 rows.

For each column:

Row 1 is the delay,

Row 2 is the number of omega parameters,

Row 3 is the number of delta parameters, and

Row 4 should take the value 1 for a simple input (rows 1, 2 and 3 are then ignored), 2 for a transfer function input for which no allowance is to be made for pre-observation period effects, and 3 for a transfer function input for which pre-observation period effects will be treated by estimation of appropriate nuisance parameters.

Note:

There will be at least one omega parameter (Omega_0) for each input series in addition to any specified by Row 2.

X_series [Optional]

The columns of array X-series must contain the original undifferenced values of each of the input series corresponding to the model specified in T_Model.

Method [Optional]

Indicates the criterion used in estimating the parameters of the model. The value 1 gives least squares, 2 gives exact likelihood, and 3 gives marginal likelihood.

Default: with no input series – exact likelihood; with input series – marginal likelihood.

Num_iterations [Optional]

The maximum required number of iterations. If Num_iterations = 0, no change is made to any of the model parameters except that the constant and any transfer function model parameters relating to simple input series are estimated.

Tolerance [Optional]

Tolerance should be set to the required convergence criterion.

Default: 0.0000001.

Output

Output as a NAGExtract format list.

Deviance

Degrees of Freedom

Variance

AIC

Parameter Estimates

Parameter Correlations

Residuals

Note:

The Parameter Estimates array consists of the parameters names, the estimates, the *t*-values for the estimates and their significance. The parameters are in the order ARMA parameters (autoregressive, moving average, seasonal autoregressive and seasonal moving average) then transfer function parameters (omega parameters followed by delta parameters) and finally the constant.

Remarks

See 'Using the Add-Ins' and the Introduction to Time Series Methods for further information.

5.3.3.7. ARIMA_FORECAST (Set 1)

ARIMA_FORECAST (NAG G13AJF) forecasts from a fitted ARIMA model.

Parameters**No_of_forecasts** [Required]

The required number of forecasts.

Series [Required]

The original, undifferenced time series.

Model [Required]

The model specification as given to ARIMA_FIT. Model must specify the orders (p, d, q, P, D, Q, s) of the (seasonal) ARIMA model to be used in the forecast.

- p: number of autoregressive terms.
- d: order of non-seasonal differencing.
- q: number of moving average terms.
- P: number of seasonal autoregressive terms.
- D: order of seasonal differencing.
- Q: number of seasonal moving average terms.
- s: the seasonality.

Parameters [Required]

The parameters for the ARIMA model specified in Model. That is, the p values for the autoregressive parameters, the q values of the moving average parameters, the P values of the seasonal autoregressive parameters, and the Q values of the seasonal moving average parameters, in that order.

Constant [Optional]

The estimated constant of the model fitted to the series.

Default: no constant in model, i.e., fixed at zero.

Output

Output as a NAGExtract format list.

Forecasts**Standard Errors of Forecasts****Remarks**

See 'Using the Add-Ins' and the Introduction to Time Series Methods for further information.

5.3.3.8. TRNS_FUNC_FORECAST (Set 1)

TRNS_FUNC_FORECAST (NAG G13BJF) produces forecasts of a time series which depends on one or more other series via a previously estimated multi-input model.

Parameters**Y_series** [Required]

The (output) time series.

Model [Required]

Model specifies the orders (p, d, q, P, D, Q, s) of the (seasonal) ARIMA model to be fitted to the output series.

- p: number of autoregressive terms.
- d: order of non-seasonal differencing.
- q: number of moving average terms.
- P: number of seasonal autoregressive terms.
- D: order of seasonal differencing.
- Q: number of seasonal moving average terms.
- s: the seasonality.

T_Model [Required]

For a multi-input (transfer function) model the *j*th column of the array T_Model specifies the form of the model relating the *j*th input series to the output series. The array must have 4 rows.

For each column:

Row 1 is the delay,

Row 2 is the number of omega parameters,

Row 3 is the number of delta parameters, and

Row 4 should take the value 1 for a simple input (rows 1, 2 and 3 are then ignored), 2 for a transfer function input for which no allowance is to be made for pre-observation period effects, and 3 for a transfer function input for which pre-observation period effects will be treated by estimation of appropriate nuisance parameters.

X_series [Required]

The columns of array X-series must contain the original undifferenced values of each of the input series corresponding to the model specified in T_Model.

Parameters [Required]

Initial values of the model parameter in the order ARMA parameters (autoregressive, moving average, seasonal autoregressive and seasonal moving average), transfer function parameters (omega parameters followed by delta parameters) and finally the constant.

Default: initial parameters are set to zero except for the constant, which is set equal to the output series mean.

X_forecasts [Required]

The columns of the array X_forecasts must contain in the forecasted values for the input series in the same order as in X-series.

Fix_constant [Optional]

Indicates whether the constant is to be fixed at its initial value or estimated.

- F: fixed at initial values.
- E: estimated.

Default: E – estimated.

X_Model [Optional]

The array X_Model may contain the ARMA model for each of the input series in the standard form of (p, d, q, P, D, Q, s). In the case of those inputs for which no ARIMA model is available, the corresponding orders should be set to 0.

RMSX [Optional]

If X_Model is set, RMSX must contain the estimated residual variance of the input series ARIMA models. In the case of those inputs for which no ARIMA model is available or its effects are not to be excluded in the calculation of forecast standard errors, the corresponding entry of RMSX should be set to 0.

ParamX [Optional]

If X_Model is set, ParamX must contain the values of the autoregressive, moving average, seasonal autoregressive and seasonal moving average parameters for each of the input series ARIMA models corresponding to the models specified in X_Model. Cells in the array that are not set to parameter values should be set to zero.

Output

Output as a NAGExtract format list.

Forecasts**Standard Errors of Forecasts****Remarks**

See 'Using the Add-Ins' and the Introduction to Time Series Methods for further information.

5.3.3.9. SPECTRAL (Set 1)

SPECTRAL (NAG G13CBF) calculates the smoothed sample spectrum of a univariate time series using spectral smoothing by the trapezium frequency (Daniell) window.

Parameters**Series** [Required]

The time series.

Width [Optional]

Determines the frequency width of the smoothing window as $2\pi/\text{Width}$. A value of n, where n denotes the length of the time-series, implies no smoothing. If present, $1 \leq \text{Width} \leq n$. Default: n – no smoothing.

Shape [Optional]

The shape parameter of the trapezium frequency window used when the width of the smoothing window has been set. A value of 0.0 gives a triangular window, and a value of 1.0 a rectangular window. If present, $0.0 \leq \text{Shape} \leq 1.0$.

Default: 1 – a rectangular window (only if Width is set).

Correction [Optional]

Whether the data are to be initially mean or trend corrected.

N: no correction,
M: mean correction,
T: trend correction.

Default: M – mean correction.

Tapering [Optional]

The proportion of the data (totalled over both ends) to be initially tapered by the split cosine bell taper. A value of 0.0 implies no tapering.

Default: 0.2

Logged [Optional]

Indicates whether unlogged or logged spectral estimates and confidence limits are required.

L: logged,
U: unlogged.

Default: U – unlogged.

Note:

If Order is set Division must also be set.

Order [Optional]

The order of the fast Fourier transform (FFT) used to calculate the spectral estimates. Order must be greater or equal to twice the length of the series. It should be a multiple of small primes such as 2^m . The largest prime factor of Order must not exceed 19 and the total number of prime factors must not exceed 20.

Default: 2^n where n represents the smallest integer such that $2^n > 2 \times \text{length of series}$.

Division [Optional]

If Order is set, the frequency division of smoothed spectral estimates as $2\pi/\text{Division}$. Division must be a factor of Order.

Default: $2^{(n-2)}$ where n represents the smallest integer such that $2^n > 2 \times \text{length of series}$ (only if Order is not set).

Output

Output as a NAGExtract format list.

Spectrum

Lower Confidence Limit

Upper Confidence Limit

Bandwidth

Degrees of Freedom

Remarks

See 'Using the Add-Ins' and the Introduction to Time Series Methods for further information.

5.4. Book 2

5.4.1. Introduction

This book contains materials for two areas:

- Analysis of Experiments
- Generalised Linear Models

Please note that a complete set of examples for this Add-In are provided in the workbook Book2E.xls. This is installed in the 'Examples' subdirectory of the Add-Ins directory. You might find it helpful to work through the examples to familiarize yourself with the use of the Book 2 Add-Ins.

5.4.2. Analysis of Experiments

5.4.2.1. Introduction

Introduction to the Analysis of Designed Experiments using the NAG Excel Add-Ins

Background

An experimental design is the plan used by the experimenter when allocating a set of controlled conditions to the experimental material. A designed experiment can be contrasted with an

observational study in which the conditions are observed as they occur naturally. The controlled conditions are known as treatments and the experimental material is divided up into units, sometimes called plots, each of which can receive a different treatment. Two examples follow.

1. In an experiment to examine the effects of different diets on the growth of chickens, the chickens were kept in pens and a different diet was fed to the birds in each pen. In this example the pens are the units and the different diets are the treatments.
2. In an experiment to compare four materials for wear-loss, a sample from each of the materials is tested in a machine that simulates wear. The machine can take four samples at a time and a number of runs are made. In this experiment the treatments are the materials and the units are the samples from the materials.

In designing an experiment the following principles are important.

Randomisation Given the overall plan of the experiment, the final allocation of treatments to units is performed using a suitable random allocation. This avoids the possibility of a systematic bias in the allocation and gives a basis for the statistical analysis of the experiment.

Replication Each treatment should be observed more than once. So in Example 2 more than one sample from each material should be tested. Replication allows for an estimate of the variability of the treatment effect to be measured.

Blocking In many situations the experimental material will not be homogeneous and there may be some form of systematic variation in the experimental material. In order to reduce the effect of systematic variation the material can be grouped into blocks so that units within a block are similar but there is variation between blocks. For example, in an animal experiment, litters may be considered as blocks; in an industrial experiment blocks may be material from one production batch.

Factorial designs If more than one type of treatment is under consideration, for example, the effect of changes in temperature and changes in pressure, a factorial design consists of looking at all combinations of temperature and pressure. The different types of treatment are known as factors and the different values of the factors that are considered in the experiment are known as levels. So if three temperatures and four different pressures were being considered, then factor 1 (temperature) would have 3 levels, factor 2 (pressure) would have four levels, and the design would be a 3×4 factorial giving a total of 12 treatment combinations. This design has the advantage of being able to detect the interaction between factors, that is, the effect of the combination of factors.

The following are examples of standard experimental designs; in the descriptions, it is assumed that there are t treatments.

1. **Completely Randomized Design** There are no blocks and the treatments are allocated to units at random.
2. **Randomized Complete Block Design** The experimental units are grouped into b blocks of t units and each treatment occurs once in each block. The treatments are allocated to units within blocks at random.
3. **Latin Square Designs** The units can be represented as cells of a square of size t classified by rows and columns. The rows and columns represent sources of variation in the experimental material. The design allocates the treatments to the units so that each treatment occurs once in each row and each column.

4. **Balanced Incomplete Block Designs** The experimental units are grouped into b blocks of t units. The treatments are allocated so that each treatment is replicated the same number of times and each treatment occurs in the same block with any other treatment the same number of times. The treatments are allocated to units within blocks at random.
5. **Complete Factorial Experiments** If there are t treatment combinations derived from the levels of all factors, then either there are no blocks or the blocks are of size t units.

Other designs include partially balanced incomplete block designs, split-plot designs, factorial designs with confounding, and fractional factorial designs. For further information on these designs, see Cochran and Cox (1957), Davies (1978) or John and Quenouille (1977).

Many of the common experimental design can be analysed using the NAG Excel Add-Ins. Three functions are available.

1. ANOVA_BLOCK handles completely Randomized designs as well as any block design with equal sized blocks. For example, Randomized block, BIBs, PBIBs can all be handled.
2. ANOVA_ROW_COLUMN handles row and column designs. For example, Latin squares and lattice squares.
3. ANOVA_FACTORIAL handles complete factorial designs that may be (unconfounded) in blocks.

Other designs can be analysed by combinations of calls to the functions. For example, a factorial experiment in a Latin square design can be analysed by first using ANOVA_ROW_COLUMN then passing the results to ANOVA_FACTORIAL for further analysis. It is also possible to analyse split-plot designs, factorial experiments in which some effects have been confounded with blocks, or some fractional factorial experiments.

The Analysis of Variance

The analysis of a designed experiment usually consists of two stages. The first is the computation of the estimate of variance of the underlying random variation in the experiment along with tests for the overall effect of treatments. This results in an analysis of variance (ANOVA) table. The second stage is a more detailed examination of the effect of different treatments either by comparing the difference in treatment means with an appropriate standard error or by the use of orthogonal contrasts.

The analysis assumes a linear model such as

$$y_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij}$$

where y_{ij} is the observed value for unit j of block i , μ is the overall mean, δ_i is the effect of the i th block, τ_j is the effect of the j th treatment which has been applied to the unit, and ε_{ij} is the random error term associated with this unit. The expected value of ε_{ij} is zero and its variance σ^2 .

In the analysis of variance, the total variation, measured by the sum of squares of observations about the overall mean, is partitioned into the sum of squares due to blocks, SS_B , the sum of squares due to treatments, SS_T , and a residual or error sum of squares, SS_E . This partition corresponds to the parameters δ , τ and σ^2 . In parallel with the partition of the sum of squares there is a partition of the degrees of freedom associated with the sums of squares. From these the mean squares are calculated as the ratio of the sum of squares divided by the degrees of freedom. Finally, a F -test for an overall effect of the treatments can be calculated as the ratio of the treatment mean square to the residual mean square. These results are usually presented in an analysis of variance (ANOVA) table as shown below.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F ratio	Significance
Blocks	b-1	SS _B	MS _B		
Treatment	t-1	SS _T	MS _T	MS _T /MS _E	
Residual	n-t-b+1	SS _E	MS _E		
Total	n-1	SS _{TOT}			

For row and column designs the model is

$$y_{ij} = \mu + \rho_i + \gamma_j + \tau_{ij}$$

where ρ_i is the effect of the i th row and γ_j is the effect of the j th column. Usually the rows and columns are orthogonal. In the analysis of variance the total variation is partitioned into rows, columns treatments and residual.

For designs in which blocks (or rows and columns) are not orthogonal to the treatments (for example in incomplete block designs) the canonical efficiency factors measure the loss of information due to the non-orthogonality (see John 1987). In such cases the treatment means computed are means adjusted for blocks.

Once the significant treatment effects have been uncovered they can be further investigated by comparing the differences between the means with the appropriate standard error. In the case of unstructured treatments these can be examined using a multiple comparison procedure or using simultaneous confidence intervals. MULT_COMPN computes simultaneous confidence intervals for the differences between means with the choice of different methods such as the Tukey-Kramer, Bonferroni and Dunn-Sidak.

When the treatments have a structure, suitable comparisons can be tested using contrasts. For example, if there are four treatments in which the first is a control and the other three are different amounts of a chemical then the contrasts that are the difference between no chemical and chemical and the linear effect of chemical could be defined. The sums of squares for these contrasts can be computed from which suitable F -tests can be performed.

ANOVA_CONTRASTS computes the sum of squares for a user-defined contrast between means. In situations where blocks and treatments are not orthogonal, both the treatment means ignoring blocks and the treatment means adjusted for blocks have to be provided.

In the case of factorial experiments, the treatment sum of squares and degrees of freedom may be partitioned into main effects for the factors and interactions between factors. The main effect of a factor is the effect of the factor averaged over all other factors. The interaction between two factors is the additional effect of the combination of the two factors, over and above the additive effects of the two factors, averaged over all other factors. For a factorial experiment in blocks with two factors, A and B, in which the i th unit of the j th block received level l of factor A and level k of factor B the model is

$$y_{ij} = \mu + \delta_j + (\alpha_l + \beta_k + \alpha\beta_{lk}) + \varepsilon_{ij}$$

where α_l and β_k are the main effects of factors A and B respectively and $\alpha\beta_{lk}$ is the interaction between level l of A and level k of B. Higher-order interactions between three or more factors can be defined in a similar way.

For factorial experiments the ANOVA table will contain a row for each of the terms. In the case of the two-factor experiment with blocks and factor A having I_A levels and factor B having I_B levels, the partition will be:

Source	Degrees of Freedom
Blocks	$b - 1$
Main effect of A	$l_a - 1$
Main effect of B	$l_b - 1$
AB interaction	$(l_a - 1)(l_b - 1)$
Residual	$n - b - l_a l_b + 1$
Total	$n - 1$

In addition to computing a table of means for each treatment term in the ANOVA table, a table of effects is also computed. The table of effects measures the additional effect of the term over and above the previous terms. For example, the AB effect measures the effect of the AB interaction in addition to the A main effect and the B main effect.

Using the ANOVA Functions

For the function ANOVA_BLOCK, the data can be presented either in the default block order or across blocks using the Block_order option. The data can also be arranged in a two-way layout and the default will be to assume that rows correspond to blocks. The treatments can be presented in full or, in the case when the data has been ordered by treatments within blocks, just the list of ordered treatments is required from which the full list can be generated. As well as the analysis of variance and treatment means with standard errors, the function also returns the canonical efficiency factor as described by John (1987).

The function ANOVA_ROW_COLUMN requires the data to be ordered by rows within columns, within replicates (if any). Again the data can be arranged in a two-way layout corresponding to rows and columns. For ANOVA_ROW_COLUMN, the full list of treatments must be provided.

Both ANOVA_BLOCK and ANOVA_ROW_COLUMN return a two-way array of covariances and standard errors for the treatments. The upper triangular part of this is the variance-covariance matrix of the treatment means, so the diagonal elements give the variances of the treatment means. The lower triangle below the diagonal contains the standard error for the differences between pairs of means. For balanced designs such as Randomized complete blocks and balanced incomplete blocks this value is constant and represents the standard error for the difference between any two treatments.

ANOVA_FACTORIAL requires the data to be ordered by the factors. That is the data is sorted by Factor 1 then by Factor 2 then by Factor 3, etc. If the data is not in the required order but the factors are available, the Excel Sort macro can be used from the 'Data' menu to re-order the data. The tables of means and effects are returned with the final factor defining the columns and the order of other factors defining the rows. For example, the A.B.C table of means would have the levels of factor C defining the columns and rows would be ordered by the levels of factor B within the levels of factor A as shown below.

Table A.B.C		Factor C	
Factor A	Factor B	1	2
1	1		
1	2		
2	1		
2	2		

All three functions have an optional parameter called `Df_adjustment`. This can be used to adjust the degrees of freedom when either analysing the data in several stages, for example using `ANOVA_ROW_COL` followed by `ANOVA_FACTORIAL`, or when estimating missing values.

Further Comments

For experiments with missing values, these values can be estimated by using the Healy and Westmacott procedure, see John and Quenouille (1977). This procedure involves starting with initial estimates for the missing values and then making adjustments based on the residuals from the analysis. The improved estimates are then used in further iterations of the process.

For designs that cannot be analysed by the above approach the function `MULT_LIN_REG` (or `NORMAL_GLM` with the Identity link) can be used to fit a general linear model.

5.4.2.2. ANOVA_BLOCK (Set 1)

`ANOVA_BLOCK` (NAG G04BBF) computes the analysis of variance for a block design with equal sized blocks or a completely Randomized design.

Parameters

No_of_Blocks [Required]

The number of blocks of equal size. For a completely Randomized design set `No_of_Blocks` to zero.

Data [Required]

The data to be analysed. The data should be in block order, i.e., Block 1 observations followed by Block 2 observations, etc., unless `Block_Order` = "A" in which case the data should be ordered across blocks, i.e., first observation from Block 1 followed by first observation from Block 2, etc.

Treatments [Optional]

The treatments applied to the Data. These would normally be present. A full set of treatments may be given but if the treatments are in a fixed order then only a list of the treatments in order is required and they will be expanded to give the full set.

Block_order [Optional]

If the data is not given by blocks, `Block_order` should be set to "A" to indicate that Data goes across blocks. When `Block_order` = "A" a list of treatments will be expanded so that the list of treatments is repeated within each block.

Tolerance [Optional]

The tolerance value used in the computations to check for zero values. If present, `Tolerance` $\geq$ 0.0.

Default: 0.00001

Df_adjustment [Optional]

This is the adjustment to the degrees of freedom for the residual and total. The degrees of freedom for the total is set to number of observations – `Df_adjustment` and the residual degrees of freedom adjusted accordingly. If present, `Df_adjustment` $\geq$ 0.

Output

The following are output when appropriate:

ANOVA table
 Treatment Means
 Covariances and Standard Errors
 Replication
 Efficiency Factors
 Grand Mean
 Block Means
 Residuals

Note:

In the Covariances and Standard Errors array, the upper triangular part contains the variance-covariance matrix of the treatment effects and the strictly lower part contains the standard errors of the difference between pairs of treatment effects.

Remarks

If no treatments are specified then the rows of the ANOVA table corresponding to Replicates, Rows, Columns and Total are computed along with the means.

See 'Using the Add-Ins' and the Introduction to the Analysis of Designed Experiments for further information.

5.4.2.3. ANOVA_ROW_COLUMN (Set 1)

ANOVA_ROW_COLUMN (NAG G04BCF) computes the analysis of variance for a general row and column design such as a Latin square.

Parameters

No_of_replicates [Required]

The number of replicates of the 'square', if only one replicate set No_of_replicates = 1.

No_of_rows [Required]

The number of rows in the 'square'.

No_of_columns [Required]

The number of columns in the 'square'.

Data [Required]

The observations ordered by columns within rows within replicates.

Treatments [Optional]

The treatments applied to the Data. These would normally be present.

Tolerance [Optional]

The tolerance value used in the computations to check for zero values. If present, Tolerance $\geq$ 0.0.

Default: 0.00001

Df_adjustment [Optional]

This is the adjustment to the degrees of freedom for the residual and total. The degrees of freedom for the total is set to number of observations – Df_adjustment and the residual degrees of freedom adjusted accordingly. If present, Df_adjustment $\geq$ 0.

Output

The following are output when appropriate:

ANOVA table
Treatment Means
Covariances and Standard Errors
Replication
Efficiency Factors
Grand Mean
Replicate Means
Row Means
Column Means
Residuals

Note:

In the Covariances and Standard Errors array, the upper triangular part contains the variance-covariance matrix of the treatment effects and the strictly lower part contains the standard errors of the difference between pairs of treatment effects.

Remarks

If no treatments are specified then the rows of the ANOVA table corresponding to Blocks and Total are computed along with the Block Means and the Grand Mean.

See 'Using the Add-Ins' and the Introduction to the Analysis of Designed Experiments for further information.

5.4.2.4. ANOVA_CONTRASTS (Set 1)

ANOVA_CONTRASTS (NAG G04DAF) computes sum of squares for a user defined contrast between treatment means.

Parameters

Treatment_means [Required]

The treatment means as computed by ANOVA_BLOCK or ANOVA_ROW_COLUMN.

Replication [Required]

The replication for each treatment mean.

rms [Required]

The residual mean square, σ^2 .

rdf [Required]

The residual degrees of freedom.

Contrasts [Required]

The columns of Contrasts contain the contrasts.

Tolerance [Optional]

The tolerance used to check if the contrasts are orthogonal and if they are orthogonal to the mean. If provided, Tolerance > 0.0.

Default: 0.0000001.

Adj_means [Optional]

If the design is non-orthogonal then the treatment means adjusted for blocks should be provided in Adj_means.

Output

The results returned consist of one row for each contrast.

ANOVA Estimates

Remarks

See 'Using the Add-Ins' and the Introduction to the Analysis of Designed Experiments for further information.

5.4.2.5. MULT_COMPN (Set 1)

MULT_COMPN (NAG G04DBF) computes simultaneous confidence intervals for the differences between means following an analysis of variance.

Parameters**Method** [Required]

Indicates which method is to be used.

- T: the Tukey–Kramer method is used.
- B: the Bonferroni method is used.
- D: the Dunn–Sidak method is used.
- L: the Fisher LSD method is used.
- S: the Scheffe method is used.

Treatment_means [Required]

The treatment means as computed by ANOVA_BLOCK and ANOVA_ROW_COLUMN.

rdf [Required]

The residual degrees of freedom from the ANOVA.

C_array [Required]

The array of covariances and standard errors as computed by ANOVA_BLOCK and ANOVA_ROW_COLUMN.

Confidence_level [Required]

The required confidence level for the computed intervals such that $0.0 < \text{Confidence_level} < 1.0$.

Output

The results are returned in upper triangular form with the first row corresponding to difference between the first treatment mean and the means of the other treatments.

Lower limits Lower limits Significance

Note:

The significance is returned as an asterisk for those comparisons that are significantly different.

Remarks

See 'Using the Add-Ins' and the Introduction to the Analysis of Designed Experiments for further information.

5.4.2.6. ANOVA_FACTORIAL (Set 1)

ANOVA_FACTORIAL (NAG G04CAF) computes an analysis of variance table and treatment means for a complete factorial design.

Parameters**No_of_blocks** [Required]

The number of blocks (replicates) in the design. If there are no blocks set No_of_blocks = 0.

Factor_names [Required]

The names of the factors in the design.

Factor_levels [Required]

The number of levels for each factor.

Data [Required]

The observations ordered by factors within blocks.

Highest_interaction [Optional]

The maximum number of factors in an interaction term. If no interaction terms are to be computed. Set Highest_interaction to 0 or 1. If present Highest_interaction ≥ 0 .

Default: all possible interaction terms computed.

Df_adjustment [Optional]

This is the adjustment to the degrees of freedom for the residual and total. The degrees of freedom for the total is set to number of observations – Df_adjustment and the residual degrees of freedom adjusted accordingly. If present, Df_adjustment ≥ 0 .

Output**ANOVA table****Tables of Treatment Means****Tables of Treatment Effects****Standard Errors of differences in means****Grand Mean****Block Means****Residuals****Note:**

The tables of means and effects consist of one array for each term and the standard errors array contains all the corresponding standard errors for the difference in pairs of treatment means.

Remarks

See 'Using the Add-Ins' and the Introduction to the Analysis of Designed Experiments for further information.

5.4.3. Generalized Linear Models

5.4.3.1. Introduction

Introduction to Generalized Linear Modelling using the NAG Excel Add-Ins

Background

Generalized Linear Models are an important extension to linear regression models that are fitted by the function MULT_LIN_REG. Examples of generalized linear models are:

1. The rational polynomial model with Normal errors: $y = 1/(a+bx+cx^2) + \varepsilon$
2. Logistic regression where y has a binomial distribution with mean μ , where $\log(\mu/(n-\mu)) = \eta = \sum \beta_i x_i$
3. Log-linear models where y can be considered as coming from a Poisson distribution with mean μ , where $\log(\mu) = \eta = \sum \beta_i x_i$

Dobson (1990) provides an introduction to generalized linear models, while further details can be found in McCullagh and Nelder (1989).

A generalized linear model consists of three components:

1. An error distribution for the response variable y .
2. A link function relating the mean, μ , of the distribution to a linear predictor, η , $\eta = g(\eta)$.
3. A linear predictor, $\sum \beta_i x_i$

For the NAG Excel Add-Ins, the following range of models is available.

Error distributions: Normal, binomial, Poisson, gamma.

Link functions:

For Normal, Poisson and gamma distributions – identity, log, exponent, square-root and reciprocal.

For Binomial distribution – logistic, probit and complementary log-log.

The linear model can be formed from the set of input columns or specified by a formula using the NAG MM_GLM function as described below.

The generalized linear model is fitted by minimizing the deviance, which is

Deviance = $-2 \times \text{Log likelihood}$ – a data dependent constant.

In the case of the Normal error this is simply the residual sum of squares. The deviance is minimized using an iterative reweighted least-squares algorithm. For further details see McCullagh and Nelder (1989) or the NAG Fortran 77 Library manual. In the case of gamma errors an adjusted deviance is used which accommodates zero values of the response variable.

In addition to the parameters of the linear model, the Normal distribution has a scale parameter, the variance. This is estimated by the residual sum of squares divided by the residual degrees of freedom.

Testing model significance

Suppose we have two nested models, for example,

$$\eta = \beta_1 + \beta_2 x$$

and

$$\eta = \beta_1 + \beta_2 x + \beta_3 x^2$$

The significance of the extra terms, in this case x^2 , can be tested by looking at the difference in the deviance between the two models. For the binomial and Poisson distributions, the difference is compared with a χ^2 distribution with degrees of freedom equal to the difference in the degrees of freedom associated with the deviances. For Normal errors, an F -statistic is computed by dividing the difference in deviance (residual sum of squares) by the difference in degrees of freedom and then dividing by the estimate of the scale parameter. This is calculated as the residual sum of squares divided by residual degrees of freedom from either the larger model or a suitable maximal model.

An individual parameter can be tested against the hypothesis that it is zero using the statistic

Parameter Estimate / Standard Error

This may be compared to a Normal distribution in the case of binomial and Poisson errors. For Normal errors the t -distribution can be used with degrees of freedom taken as the degrees of freedom associated with the deviance or residual sum of squares.

Model checking

In addition to testing the model parameters, the residuals from the model should be examined. For Normal, binomial and Poisson errors, the deviance residuals are computed. In the case of the Normal errors the deviance residual is simply $y - \mu$. For gamma errors, Anscombe residuals that are defined for $y = 0$ are used.

In addition to residuals, leverages are computed. These show the influence of the observation on the fitted value. A large value indicates that the fitted model may be unduly influenced by that observation. Leverages can also be used to standardize residuals. See STAND_RESIDS for more information.

For advanced use, the generalized linear model functions also return the linear predictor and the working weights used in the iterative re-weighted least squares. The variance standardization is also returned for the binomial, Poisson and gamma distributions.

Specifying the Model Matrix

A simple term selection mechanism is given by the optional Select_x parameter of the generalized linear model functions. This allows columns from the X array to be excluded from the model. A more powerful model specification is provided by the NAG MM_GLM function.

The NAG function MM_GLM can be used to specify a model matrix rather than the user constructing a set of columns. The inputs to the model matrix function MM_GLM are

1. Group variables with their names
2. Covariates with their names
3. A formula

Group variables are classification variables, e.g., eye colour, sex. These variables can be either text, e.g., red, green, blue, or numerical, 100, 150, 300. MM_GLM re-codes these values first into an integer representation, 1,2,..., then into the dummy variables required for the model fitting. Covariates are ordinary variables, e.g., weight, temperature. A formula uses the variable

names and the symbols

$+ - * / () ^$

Some examples of formulae are:

Group variables: a, b and c

Covariates: x and z

$a + x$ fits the group variable a and covariate x, that is, regression model with different intercepts for the different groups.

$x + x^2$ fits the quadratic model in x.

$a*b$ fits the terms $a + b + a.b$, that is, the main effects of a and b and the a.b interaction, that is, a cross-classification model.

a/b fits $a + a.b$, that is the nested model.

$(a + b)*c$ fits $a*c$ plus $b*c$, that is, $a + b + c + a.c + b.c$

The ^ notation is only available for use with covariates. There is also an additional parameter, Interaction, which sets the maximum number of terms allowed in an interaction. So, for example,

$a*b*c$ with Interaction = 2 gives $a + b + c + a.b + a.c + b.c$ with the three term interaction excluded.

Note that the model matrix produced by MM_MLR assumes that the mean is also fitted. If the mean is to be omitted then DUMMY_VARS should be used to construct the required dummy variables with all levels present.

5.4.3.2. NORMAL_GLM (Set 1)

NORMAL_GLM (NAG G02GAF) fits a generalized linear model with normal errors.

Parameters

Link [Required]

Indicates which link function is to be used.

E: an exponent link is used.

I: an identity link is used.

L: a log link is used.

S: a square root link is used.

R: a reciprocal link is used.

Y [Required]

The response variable.

X [Required]

Either the array containing the independent variables as columns or the function MM_GLM that computes the model matrix.

Select_x [Optional]

Indicates which independent variables in array X are to be included in the model. If $\text{Select_x}(j) > 0$, then the variable contained in the j th column of array X is included in the model, if $\text{Select_x}(j) = 0$ it is excluded.

If Select_x is present then its elements ≥ 0 .

Default: all X variables are included.

X_names [Optional]

Names for the explanatory variables.

Default: X_1, etc.

Offsets [Optional]

Offsets for the model.

Default: none.

Weights [Optional]

User weights. If a Weight is set to zero the observation is not included in the model, in which case the effective number of observations is the number of observations with non-zero weights. If present, $\text{Weights} \geq 0.0$.

Default: no weights.

Mean [Optional]

Indicates if a mean term is to be included.

- M: a mean term, intercept, is to be included in the model.
- Z: the model will pass through the origin, zero point.

Default: mean term included.

Power [Optional]

If an exponent link is selected, the power of the exponent.

Restriction: $\text{Power} \neq 0.0$.

S [Optional]

The scale parameter for the model, σ^2 . If omitted, the scale parameter is estimated by the residual mean square. If present, $s \geq 0.0$.

Fitted_values [Optional]

Indicates if fitted values etc. are to be returned.

- N: No fitted values etc. are returned.
- Y: Fitted values and residuals are returned.
- A: All of fitted values, residuals, leverages, linear predictor, variance function and working weights are returned.

Default: Y – Fitted values and residuals are returned

Max_iterations [Optional]

The maximum number of iterations for the iterative weighted least-squares fitting.

If present, $\text{Max_iterations} > 0$.

Default: 10

Tolerance [Optional]

Indicates the accuracy required for the fit of the model. If present, Tolerance > 0.0.

Default: 0.000001

Eps [Optional]

The value of eps is used to decide if the independent variables are of full rank, and, if not, the rank of the independent variables. If present, eps > 0.0.

Default: 0.00000000001

Output

Output as a NAGExtract format list.

RSS**df****Rank of Model****Parameter Estimates****Cov**

If requested (see Fitted_values) the following are returned.

Fitted values**Residuals****Leverages****Linear predictor****Working weights****Notes:**

The residual sum of squares (RSS) and its degrees of freedom (df) are returned providing the residual degrees of freedom are not zero. The Rank of Model is returned only when the model is not of full rank.

The Parameter Estimates consist of the four columns with headings giving:

Parameter Names
Parameter Estimate
Standard Error
t-value

The variance-covariance matrix of the parameters (Cov) is returned as an upper triangular matrix.

Remarks

See 'Using the Add-Ins' and the Introduction to Generalized Linear Models for further information.

5.4.3.3. BINOMIAL_GLM (Set 1)

BINOMIAL_GLM (NAG G02GBF) fits a generalized linear model with binomial errors.

Parameters**Link** [Required]

Indicates which link function is to be used.

G: a logistic link is used.
P: a probit link is used.
C: a complementary log-log link is used.

Y [Required]

The response variable.

Denominator [Required]

The binomial denominator.

X [Required]

Either the array containing the independent variables as columns or the function MM_GLM that computes the model matrix.

Select_x [Optional]

Indicates which independent variables in array X are to be included in the model. If $\text{Select_x}(j) > 0$, then the variable contained in the j th column of array X is included in the model, if $\text{Select_x}(j) = 0$ it is excluded. If Select_x is present then its elements ≥ 0 .

Default: all X variables are included.

X_names [Optional]

Names for the explanatory variables.

Default: X_1, etc.

Offsets [Optional]

Offsets for model.

Default: none.

Weights [Optional]

User weights. If a weights is set to zero, the observation is not included in the model, in which case the effective number of observations is the number of observations with non-zero weights. If present, $\text{Weights} \geq 0.0$.

Default: no weights.

Mean [Optional]

Indicates if a mean term is to be included.

M: a mean term, intercept, is to be included in the model.

Z: the model will pass through the origin, zero point.

Default: mean term included.

Fitted_values [Optional]

Indicates if fitted values etc. are to be returned.

N: No fitted values etc. are returned.

Y: Fitted values and residuals are returned.

A: All of fitted values, residuals, leverages, linear predictor, variance function and working weights are returned.

Default: Y – Fitted values and residuals are returned.

Max_iterations [Optional]

The maximum number of iterations for the iterative weighted least-squares fitting. If present $\text{Max_iterations} > 0$.

Default: 10

Tolerance [Optional]

Indicates the accuracy required for the fit of the model. If present, Tolerance > 0.0.

Default: 0.000001

Eps [Optional]

The value of eps is used to decide if the independent variables are of full rank, and, if not, the rank of the independent variables. If present, eps > 0.0.

Default: 0.00000000001

Output

Output as a NAGExtract format list.

Deviance**df****Rank of Model****Parameter Estimates****Cov**

If requested (see Fitted_values) the following are returned.

Fitted values**Residuals****Leverages****Linear predictor****Variances****Working weights****Notes:**

The deviance and its degrees of freedom (df) are returned providing the degrees of freedom are not zero. The Rank of Model is returned only when the model is not of full rank.

The Parameter Estimates consist of the four columns with headings giving:

Parameter Names
Parameter Estimate
Standard Error
t-value

The variance-covariance matrix of the parameters (Cov) is returned as an upper triangular matrix.

Remarks

See 'Using the Add-Ins' and the Introduction to Generalized Linear Models for further information.

5.4.3.4. POISSON_GLM (Set 1)

POISSON_GLM (NAG G02GCF) fits a generalized linear model with Poisson errors.

Parameters**Link** [Required]

Indicates which link function is to be used.

- E: an exponent link is used and the power must be provided.
- I: an identity link is used.
- L: a log link is used.
- S: a square root link is used.
- R: a reciprocal link is used.

Y [Required]

The response variable.

X [Required]

Either the array containing the independent variables as columns or the function MM_GLM that computes the model matrix.

Select_x [Optional]

Indicates which independent variables in array X are to be included in the model. If $\text{Select_x}(j) > 0$, then the variable contained in the j th column of array X is included in the model, if $\text{Select_x}(j) = 0$ it is excluded.

If Select_x is present then its elements ≥ 0 .

Default: all variables included.

X_names [Optional]

Names for the explanatory variables

Default: X_1, etc.

Offsets [Optional]

Offsets for model.

Default: none.

Weights [Optional]

User weights. If a weight is set to zero the observation is not included in the model, in which case the effective number of observations is the number of observations with non-zero weights.

If present, $\text{Weights} \geq 0.0$.

Default: no weights.

Mean [Optional]

Indicates if a mean term is to be included.

- M: a mean term, intercept, is to be included in the model.
- Z: the model will pass through the origin, zero point.

Default: M – mean term included.

Power [Optional]

If an exponent link is selected, then the power of the exponent. If present, $\text{Power} \neq 0.0$.

Fitted_values [Optional]

Indicates if fitted values etc. are to be returned.

- N: No fitted values etc. are returned.
- Y: Fitted values and residuals are returned.
- A: All of fitted values, residuals, leverages, linear predictor, variance function and working weights are returned.

Default: Y – Fitted values and residuals are returned.

Max_iterations [Optional]

The maximum number of iterations for the iterative weighted least-squares fitting. If present Max_iterations > 0.

Default: 10

Tolerance [Optional]

Indicates the accuracy required for the fit of the model. If present, Tolerance > 0.0.

Default: 0.000001

Eps [Optional]

The value of eps is used to decide if the independent variables are of full rank, and, if not, the rank of the independent variables. If present, eps > 0.0.

Default: 0.00000000001

Output

Output as a NAGExtract format list.

Deviance**df****Rank of Model****Parameter Estimates****Cov**

If requested (see Fitted_values) the following are returned.

Fitted values**Residuals****Leverages****Linear predictor****Variances****Working weights**

Notes:

The deviance and its degrees of freedom (df) are returned providing the degrees of freedom are not zero. The Rank of Model is returned only when the model is not of full rank.

The Parameter Estimates consist of the four columns with headings giving:

- Parameter Names
- Parameter Estimate
- Standard Error
- t-value

The variance-covariance matrix of the parameters (Cov) is returned as an upper triangular matrix.

Remarks

See 'Using the Add-Ins' and the Introduction to Generalized Linear Models for further information.

5.4.3.5. GAMMA_GLM (Set 1)

GAMMA_GLM (NAG G02GDF) fits a generalized linear model with gamma errors.

Parameters**Link** [Required]

Indicates which link function is to be used.

- E: an exponent link is used.
- I: an identity link is used.
- L: a log link is used.
- S: a square root link is used.
- R: a reciprocal link is used.

Y [Required]

The response variable.

X [Required]

Either the array containing the independent variables as columns or the function MM_GLM that computes the model matrix.

Select_x [Optional]

Indicates which independent variables in array X are to be included in the model. If $\text{Select_x}(j) > 0$, then the variable contained in the j th column of array X is included in the model, if $\text{Select_x}(j) = 0$ it is excluded.

If Select_x is present then its elements ≥ 0 .

Default: all variables included.

X_names [Optional]

Names for the explanatory variables.

Default: X_1, etc.

Offsets [Optional]

Offsets for the model.

Default: none.

Weights(n) [Optional]

User weights. If a Weight is set to zero observation is not included in the model, in which case the effective number of observations is the number of observations with non-zero weights. If present, $\text{Weights} \geq 0.0$.

Default: no weights.

Mean [Optional]

Indicates if a mean term is to be included.

- M: a mean term, intercept, is to be included in the model.
- Z: the model will pass through the origin, zero point.

Default: mean term included

Power [Optional]

If an exponent link is selected, the power of the exponent. If present, Power $\neq$ 0.0.

S [Optional]

The scale parameter for the model, σ^2 . If omitted, the scale parameter is estimated by the residual mean square. If present, $s \geq 0.0$.

Fitted_values [Optional]

Indicates if fitted values etc. are to be returned.

- N: No fitted values etc. are returned
- Y: Fitted values and residuals are returned
- A: Fitted values, residuals, leverages, linear predictor, variance function and working weights are returned

Default: Y – Fitted values and residuals are returned

Max_iterations [Optional]

The maximum number of iterations for the iterative weighted least-squares fitting. If present, Max_iterations > 0.

Default: 10

Tolerance [Optional]

Indicates the accuracy required for the fit of the model. If present, Tolerance > 0.0.

Default: 0.000001

Eps [Optional]

The value of eps is used to decide if the independent variables are of full rank, and, if not, the rank of the independent variables. If present, eps > 0.0.

Default: 0.00000000001

Output

Output as a NAGExtract format list.

Deviance

df

Rank of Model

Parameter Estimates

Cov

If requested (see Fitted_values) the following are returned.

Fitted values

Residuals

Leverages

Linear predictor

Variances

Working weights

Notes:

The adjusted deviance and its degrees of freedom (df) are returned providing the degrees of freedom are not zero. The Rank of Model is returned only when the model is not of full rank.

The Parameter Estimates consist of the four columns with headings giving:

Parameter Names
Parameter Estimate
Standard Error
t-value

The variance-covariance matrix of the parameters (Cov) is returned as an upper triangular matrix.

Remarks

See 'Using the Add-Ins' and the Introduction to Generalized Linear Models for further information.

5.4.3.6. MM_GLM (Set 1)

MM_GLM computes the model matrix for a linear model specified by a formula. It is for use with NORMAL_GLM, BINOMIAL_GLM, POISSON_GLM and GAMMA_GLM where it is called instead of providing the X array.

Parameters

Formula [Required]

The model formula that may consists of the names of the variables given in Group_Names and Covariate_Names and the following symbols:

() + . * / – ^

where

- + add term to formula
- . interaction between two variables
- * cross-classify two terms, e.g., $a*b = a + b + a.b$
- / nest two terms, e.g. $a/b = a + a.b$
- remove term from formula, e.g., $a*b - a.b = a + b$
- ^ power for covariates only

The precedence of the operators is () – + * / . ^

At least one of Groups or Covariates must be given along with the corresponding names.

Groups [Optional]

The array whose columns contain the variables that are to be considered as grouping or classification variables.

Group_Names [Optional]

The names of the variables in Groups.

Covariates [Optional]

The array whose columns contain the variables that are to be considered as covariates, i.e., not as classification variables.

Covariate_Names [Optional]

The names of the variables in Covariates.

Interaction [Optional]

The maximum number of terms to be included in an interaction, ≥ 1 .

Default: all terms included.

Output

To avoid the restriction on the size of returned arrays the function MM_GLM passes the results to the calling function by means of global variables. All that is returned directly is an indicator string or any error messages.

The function computes all required dummy variables with appropriate names. The dummy variable corresponding to the first level of a group variable is not included in order to provide a full rank model.

Remarks

See 'Using the Add-Ins' and the Introduction to Generalized Linear Models for further information.

5.5. Book 3

5.5.1. Introduction

Introduction to Multivariate Methods

A multivariate data set consists of several variables recorded on a number of objects or individuals. For example, the amounts of different minerals present in a number of soil samples or different social indicators for regions of a city. Multivariate methods can be classified as those that seek to examine the relationships between the variables (e.g., principal components), known as variable-directed methods, and those that seek to examine the relationships between the objects (e.g., cluster analysis), known as individual-directed methods.

Please note that a complete set of examples for this Add-In are provided in the workbook Book3E.xls. This is installed in the 'Examples' subdirectory of the Add-Ins directory. You might find it helpful to work through the examples to familiarize yourself with the use of the Book 3 Add-Ins.

5.5.2. Variable-Directed Methods

5.5.2.1. Introduction

Variable-directed methods

Most variable-directed methods seek to find derived variables that represent linear relationships between the variables with the aim of reducing the dimensionality of the problem by considering only a small number of these derived variables. The two most common methods are principal component analysis and factor analysis.

Principal component analysis finds new variables that are linear combinations of the observed variables so that they have maximum variation and are orthogonal (uncorrelated). The required variables can be shown to be the eigenvectors of the sample variance-covariance matrix. The amount of variation explained by the eigenvector is proportional to its corresponding eigenvalue. The eigenvector corresponding to the largest eigenvalue is known as the first principal component. As an alternative to the variance-covariance matrix the correlation matrix can be used. This has the effect of standardizing the variables to the same scale. Ideally only a small

number of dimensions (principal components) are needed to represent most of the variation in the data. A test on the smaller eigenvalues can be used to investigate the number of dimensions needed. The coefficients provided by the eigenvalues indicate how the components relate to the original variables.

The values of the principal component variables for the individuals are known as the principal component scores. These can be standardized so that the variance or the sum of squares of these scores for each principal component is equal to 1.0 or the corresponding eigenvalue.

The function PRIN_COMP computes the principal component analysis and returns the eigenvalues, the proportion of the variation explained by each component and tests for the number of components required. The coefficients of the eigenvectors and the scores are also returned.

Factor analysis is similar to principal component analysis but uses a different underlying model. Like principal component analysis, the aim of factor analysis is to account for the covariances in these variables in terms of a smaller number of latent variables, or factors. These variables are assumed to be independent and to have unit variance. In factor analysis the relationship between the variance-covariance of the observed variables, S , and the factors is given by the model

$$S = LL' + \psi$$

where L is the matrix of the factor loadings, and ψ is a diagonal matrix of the unique variances for each variable. The values $1 - \psi_i$ are known as the communalities. The loadings relate the observed variables, Y , to the unobserved factors, F . $Y = LF + e$. If it is assumed that both the factors and the unique components, e , follow independent Normal distributions then the parameters of the model, that is, L and ψ , can be estimated by maximum likelihood as described by Lawley and Maxwell (1971). The computation of the maximum likelihood estimates is an iterative procedure that involves computing the eigenvalues and eigenvectors of the matrix

$$S^* = \psi^{-1/2} S \psi^{-1/2}$$

where S is the sample variance-covariance matrix. The use of maximum likelihood estimation means that likelihood ratio tests can be constructed to test for the number of factors required by comparing log-likelihoods. Having found the estimates of the parameters of the model, the estimates of the values of the factors for the individuals, the factor scores, can be computed. These involve the calculation of the factor score coefficients. Two common methods of computing factor score coefficients are the regression method and Bartlett's method. Bartlett's method gives unbiased estimates of the factor scores whereas estimates from the regression method are biased but have smaller variance than those from Bartlett's method; see Lawley and Maxwell (1971). The function FACTOR computes the loadings and communalities. Either the raw data or the correlation/variance-covariance matrix can be input. The function FACTOR_SCORE can then be used to compute the factor score coefficients. To compute the factor scores from the coefficients the data has to be standardized and then post-multiplied by the factor score coefficients.

With both principal component analysis and factor analysis the pattern of the coefficients or loadings is used as a basis of interpreting the components or factors. The interpretation is easier if the coefficients are either very small or large in absolute value. This can often be achieved by using an orthogonal rotation. Two methods are commonly used: varimax and quartimax. In both cases the rotation is chosen to maximize a criterion that involves the fourth power of the coefficients; for further details see, for example, Lawley and Maxwell (1971) or Krzanowski (1990). The function ORTHOG_ROTATION computes the orthogonal rotation. If the parameter G is set to unity then varimax rotation is performed and if G is set to zero

quartimax rotation is performed. Values of g between zero and one would give a compromise between the two methods.

5.5.2.2. PRIN_COMP (Set 1)

PRIN_COMP (NAG G03AAF) performs a principal component analysis on a data matrix.

Parameters

Data [Required]

The array of observed variables in columns.

Matrix_type [Required]

Indicates for which type of matrix the principal component analysis is to be carried out.

- C: for the correlation matrix.
- S: for a matrix standardized by values in User_std.
- U: for the sums of squares and cross-products matrix.
- V: for the variance-covariance matrix.

Std_scores [Required]

Indicates if the principal component scores are to be standardized.

- S: they are standardized to have sum of squares equal to one.
- U: they are unstandardized, and have sum of squares equal to the corresponding eigenvalue.
- Z: they are standardized so that they have unit variance.
- E: they are standardized so that they have variance equal to the corresponding eigenvalue.

Number_of_scores [Optional]

The number of scores to be returned.

User_std [Optional]

The standardizations to be used for each variable if Matrix_type = S. The mean corrected data values are divided by these values. If User_std is set equal to the standard deviations it would be equivalent to analysing a correlation matrix (Matrix_type = C). If present, values must be > 0 .

Weights [Optional]

Contains the weights corresponding to the observations. If $\text{Weights}(i) = 0.0$ then the i th observation is not included in the analysis. If present, values must be ≥ 0.0

Output

Output as a NAGExtract format list.

Eigenvalues Coefficients Scores

Notes:

The Eigenvalues table consists of the eigenvalues, the proportion explained by each component and the cumulative proportion. In addition, except for when the correlation matrix is analysed, test value, its degrees of freedom and its significance are returned.

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.2.3. FACTOR (Set 1)

FACTOR (NAG G03CAF) computes the maximum likelihood estimates of the parameters of a factor analysis model. Either the data matrix or a correlation/covariance matrix may be input.

Parameters

No_of_factors [Required]

The number of factors that are required

Note: Give EITHER Correlation_matrix and Number_of_obs OR Data_matrix and Matrix_type (if default not used).

Correlation_matrix [Optional].

The correlation or variance-covariance matrix; only the upper triangular part is required.

Number_of_obs [Optional]

The number of observations used in computing the correlation/variance-covariance matrix.

Data_matrix [Optional]

The data matrix with the variables in columns.

Matrix_type [Optional]

Selects the type of matrix on which factor analysis is to be performed.

C: the factor analysis will be computed for the correlation matrix.

V: the factor analysis will be computed for the covariance matrix.

Default: C – Correlation matrix when data matrix supplied

Weights [Optional]

If required, the weights for each the observations. If $\text{Weights}(i) = 0.0$ then the i th observation is not included in the analysis. If present, $\text{Weights} \geq 0.0$.

Max_steps [Optional]

The maximum number of function evaluations used in computing the estimates.

Default: 100 times number of variables.

Acc_req [Optional]

The number of decimal places of accuracy required for the estimates of the unique component of the variation, y . If present, $0 < \text{Acc_req} < 15$.

Default: 7

Lower_phi [Optional]

The number of decimal places for the lower bound for the values of y , i.e., the lower bound is $10^{-\text{Lower_phi}}$. If present, $0 < \text{Lower_phi} < 15$.

Default: 13

Output

Output as a NAGExtract format list.

Eigenvalues

Communalities

Psi

Loadings
Function value
Test statistic
Degrees of freedom
Significance
Residuals (as lower triangular)

Note:

The Test statistic, Degrees of freedom and Significance are the value of the chi-squared statistic for testing the goodness of fit of the model along with its degrees of freedom and significance.

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.2.4. FACTOR_SCORE (Set 1)

FACTOR_SCORE (NAG G03CCF) computes factor score coefficients from the result of fitting a factor analysis model by maximum likelihood as performed by FACTOR.

Parameters

Method [Required]

Indicates which method is to be used to compute the factor score loadings.

R: the regression method is used.

B: Bartlett's method is used.

Loadings [Required]

The matrix of unrotated factor loadings.

Psi [Required]

The values of the unique component of the variation, ψ , as returned by FACTOR.

Eigenvalues [Required]

The eigenvalues as returned by FACTOR.

Rotation [Optional]

If rotations have been used, the orthogonal rotation matrix as returned by ORTHOG_ROTATION.

Output

Output as a NAGExtract format list

Factor Score Coefficients

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.2.5. ORTHOG_ROTATION (Set 1)

ORTHOG_ROTATION (NAG G03BAF) computes orthogonal rotations for a matrix of loadings.

Parameters

Loadings [Required]

The array of loadings, one column per factor.

G [Optional]

G = 1.0 gives varimax rotations and G = 0.0 giving quartimax rotations. If present, $g \geq 0.0$.

Default: G = 1.0, i.e., varimax rotations.

Stand_rows [Optional]

Indicates if the matrix of loadings is to be row standardized before rotation.

S: loadings are row standardized.

U: loadings are unstandardized.

Default: S – standardized

Acc_req [Optional]

The accuracy required. If present, $\text{Acc_req} \geq 0.0$.

Default: 0.00001

Max_steps [Optional]

The maximum number of iterations required. If present, $\text{Max_steps} \geq 1$.

Default: 30

Output

Output as a NAGExtract format list.

Rotated Loadings**Rotation Matrix****Remarks**

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.3. Individual-Directed Methods

5.5.3.1. Introduction

Individual-directed Methods

In general the derived variables will be unique only up to an orthogonal rotation. Therefore it may be useful to see if suitable rotations of these variables exist that lead to a simple interpretation of the new variables in terms of the original variables.

While dealing with the same data matrix as variable-directed methods the emphasis is the objects or individuals rather than the variables. The methods are generally based on distance or dissimilarity matrices such that the elements give a measure of how 'far apart' individuals are. Alternatively, a similarity matrix can be used which measures how 'close' individuals are. The form of the measure of distance or similarity will depend upon the form of the variables. For continuous variables it is usually assumed that some form of (squared) Euclidean distance is suitable.

$$D_{xy} = (x - y)^2.$$

Another commonly used distance is the absolute distance or city block metric:

$$D_{xy} = |x - y|.$$

The function DIST_MAT can be used to compute these distances.

Often there will be a need to scale the variables to produce satisfactory distances. For discrete variables there are various measures of similarity or distance that can easily be computed. For example, for binary data a measure of dissimilarity could be

- 0 if the individuals take the same value,
- 1 otherwise.

As DIST_MAT has an update option, distances computed in this way can be combined with the Euclidean distances given above.

Given a measure of distance between individuals there are three basic tasks that can be performed.

1. **Group the individuals:** that is, collect the individuals into groups so that those within a group are closer to each other than they are to members of another group.
2. **Classify individuals:** that is, if some individuals are known to come from certain groups, allocate individuals whose group membership is unknown to the nearest group.
3. **Map the individuals:** that is, produce a multidimensional diagram in which the distances on the diagram represent the distances between the individuals.

In the above, (1) leads to cluster analysis, (2) leads to discriminant analysis, and (3) leads to scaling methods.

Hierarchical cluster analysis

Approaches for cluster analysis can be classified into two types: hierarchical and non-hierarchical. Hierarchical cluster analysis produces a series of overlapping groups or clusters ranging from separate individuals to one single cluster. This forms a tree. The clusters at a level are constructed from the clusters at a previous level. There are two basic approaches to hierarchical cluster analysis: agglomerative methods, which build up clusters starting from individuals until there is only one cluster, or divisive methods, which start with a single cluster and split clusters until the individual level is reached. This Book contains the more common agglomerative methods.

The stages in a hierarchical cluster analysis are usually as follows.

1. Form a distance matrix.
2. Use the selected criterion to form the hierarchy.
3. Print the cluster information in the form of a dendrogram or use the information to form a set of clusters.

Given a distance matrix for the n individuals as described above, an agglomerative clustering method produces a hierarchical tree by starting with n clusters each with a single individual and then at each of $n - 1$ stages merging two clusters to form a larger cluster until all individuals are in a single cluster, thus forming a tree. At each stage the two clusters that are nearest are merged to form a new cluster and a new distance matrix is computed for the reduced number of clusters.

The methods differ as to how the distances between the new cluster and other clusters are computed. Six methods are available:

1. Single link or nearest neighbour
2. Complete link or furthest neighbour
3. Group average
4. Centroid
5. Median
6. Minimum variance

Let there be three clusters A, B and C with distances

	A	B	C
A		1.0	2.0
B			3.0
C			

At the next step cluster A is merged with cluster B. The distances between the new cluster and cluster C will be 2 (the minimum distance from A or B to C) if single link is chosen or 3 (the maximum distance from A or B to C) if complete link is chosen. Group average takes into account the size of the clusters A and B. So if cluster A had six objects and cluster B four then the new distance to C would be

$$2(6/10) + 3(4/10) = 2.4.$$

The other three methods also take into account the distance between A and B. Median clustering takes the average of the distances minus a quarter of the AB distance, e.g.,

$$(2+3)/2 - \frac{1}{4} = 2.25.$$

Centroid clustering adjusts the group average distance by the AB distance giving

$$2(6/10) + 3(4/10) - (6/10)(4/10) = 2.16.$$

Finally, minimum variance also takes into account the number of objects in C. If there were 5 objects in C, minimum variance clustering would give the distance as

$$((5+6)2 + (5+4)3 - (5)1)/(5+6+4) = 2.93.$$

Details of these methods can be found in Everitt (1974) or Krzanowski (1990).

Once the tree has been formed, the clusters can be created by either specifying the number of clusters required or specifying the distance at which clusters are formed.

The function CLUSTER computes a hierarchical cluster analysis tree using one of the six methods described above. You can enter either the distance matrix or enter the raw data and use one of the Euclidean distances. The function GROUPS_FROM_CLUSTER can then be used to compute the required clusters.

Discriminant analysis

Discriminant analysis is concerned with the allocation of objects to groups on the basis of observations on those objects using an allocation rule. This rule is computed from observations coming from a training set in which group membership is known. The allocation rule is based on the distance between the object and an estimate of the location of the groups. The usual measure of the distance of an observation from group mean is given by Mahalanobis distance in which the squared distances are standardized by the variance-covariance matrix of the variables. This can either be the variance-covariance matrix for the group or a pooled variance-covariance matrix if the group variance-covariance matrices can be assumed to be equal. This

assumption can be tested; see Morrison (1967). In addition to the distances a set of prior probabilities of group may be used. The prior probabilities reflect the user's view as to the likelihood of the objects coming from the different groups.

It is generally assumed that the variables follow a multivariate Normal distribution. By combining the prior probability of an observation's membership of a group with the probability of the observation given a group, a posterior probability of belonging to a group can be computed. An observation is then allocated to the group with the highest posterior probability.

In the estimative approach to discrimination, the parameters of the multivariate Normal distribution which appear in the posterior are replaced by their estimates calculated from the training set. If it is assumed that the within-group variance-covariance matrices are equal then the linear discriminant function is obtained; otherwise if it is assumed that the variance-covariance matrices are unequal then the quadratic discriminant function is obtained. In the Bayesian predictive approach, a non-informative prior distribution is used for the parameters giving the posterior distribution for the parameters from the training set. A predictive distribution is then obtained by integrating over the parameter space. This predictive distribution is then used in place of the posterior distribution. In addition to allocating the objects to groups, an atypicality index for each object and for each group can be computed. This represents the probability of obtaining an observation more typical of the group than that observed. A high value of the atypicality index for all groups indicates that the observation may in fact come from a group not represented in the training set.

The function DISCRIM_TEST computes a test statistic for the equality of the group variances and also computes the statistics required to carry out a discriminant analysis. The information from DISCRIM_TEST can be passed to DISCRIM_ALLOC that will then allocate observations to groups on the basis of this information. The choice of pooled or group variances is available, as is the choice between the estimative and the predictive approach. Posterior/predictive probabilities as well as atypicality indices are computed. The function DISCRIM_DIST can be used to compute Mahalanobis distances after the use of DISCRIM_TEST. Either the between-group means or between observations and group means are produced. An alternative approach to discrimination is the logistic discrimination. This does not depend on the assumption that the data has a Normal distribution. In the case of two groups, logistic regression can be performed with the response variable indicating the group allocation and the variables in the discriminant analysis being the explanatory variables. Allocation can then be made on the basis of the fitted response value. This approach can be shown to be valid for a wide range of distributional assumptions. The function BINOMIAL_GLM is available in Book 2.

Scaling methods

Scaling methods seek to represent the observed dissimilarities or distances between objects as distances between points in Euclidean space. For example, if the distances between objects A, B and C were 3, 4 and 5 respectively the distances could be represented exactly by three points in two-dimensional space. Only their relative positions would be important; the whole configuration of points could be rotated or shifted without affecting the distances between the points. If a one-dimensional representation was required, the 'best' representation might give distances of $7/3$, $10/3$ and $17/3$ which may be an adequate representation. If the distances were 3, 4 and 8 then these distances could not be exactly represented in Euclidean space, even in two dimensions, the best representation being the three points in a straight line, yielding distances 3, 4 and 7.

In practice the user of scaling methods has to decide upon the number of dimensions in which the data is to be represented. The smaller the number the easier it will be to assimilate the information. The chosen number of dimensions needs to give an adequate representation of the data but will often not give an exact representation because either the number of chosen dimensions is too small or the data cannot be represented in Euclidean space.

Two basic methods are available depending on the nature of the dissimilarities or distances being analysed. If the distances can be assumed to satisfy the metric inequality

$$d_{ij} \geq d_{ik} + d_{kj},$$

then the distances can be represented exactly by points in Euclidean space and the technique known as metric scaling, classical scaling or principal co-ordinate analysis can be used. This technique involves the computing of the eigenvalues of a matrix derived from the distance matrix. The eigenvectors corresponding to the k largest positive eigenvalues give the best k dimensions in which to represent the objects. The function METRIC_SCAL performs this type of scaling. If there are negative eigenvalues then the distance matrix cannot be represented in Euclidean space. METRIC_SCAL gives the option of computing all the eigenvalues so that this assumption can be checked.

Instead of the above approach (of requiring the distances from the points to match the distances from the objects as closely as possible) sometimes only a rank-order equivalence is required. That is, the order of the distances between the points should, as far as possible, be the same as the distances between the objects they represent. This would be appropriate when the dissimilarities are based on subjective rankings. For example, if the objects were foods and a number of judges ranked the foods for different qualities such as taste and texture, the resulting distances would not necessarily obey the metric inequality but the rank order would be significant. Alternatively, by relaxing the requirement from matching distances to rank order equivalence only, the number of dimensions required to represent the distance matrix may be decreased. The requirement of rank-order equivalence leads to non-metric or ordinal multidimensional scaling. The criterion used to measure the closeness of the fitted distance matrix to the observed distance matrix is known as STRESS. Non-metric multidimensional scaling seeks to find the set of points that minimize the STRESS.

STRESS is measured by how well the set of points preserve the order of the distances in the original distance matrix. The computation of STRESS involves the monotonic regression of the fitted distances on the observed distances. For example

Observed Distance	1.05	1.25	1.75	2.6
Fitted Distance	0.93	2.62	1.21	3.55
Monotonic Distance	0.93	1.85	1.85	3.55

The STRESS is then computed from the sum of squared differences between the fitted distance and the monotonic distance divided by the sum of squared fitted distances. As an alternative to STRESS, squared STRESS, SSTRESS, can be used in which squared distances are preferred instead of the distances themselves. The function MDS performs non-metric scaling with a choice of STRESS or squared STRESS. In addition to the points, MDS has the option of returning the fitted distances and the monotonic distances. These can be either as a distance matrix corresponding to the input distance matrix or as arrays ordered by the observed distances. The latter can be used in plots to assess the appropriateness of the method. MDS requires the input of an initial configuration of the points. This can usually be provided by METRIC_SCAL. Unlike METRIC_SCAL, MDS allows missing values in the distance matrix to be specified as negative distances; the algorithm may still work provided there are not more than two thirds of the values missing.

5.5.3.2. DIST_MAT (Set 1)

DIST_MAT (NAG G03EAF) computes a distance (dissimilarity) matrix.

Parameters**Distance** [Required]

Indicates which type of distances are computed.

- A: absolute distances.
- E: Euclidean distances.
- S: Euclidean squared distances.

Scaling [Required]

Indicates the standardization of the variables to be used.

- S: standard deviation.
- R: range.
- G: standardizations given in array S.
- U: unscaled.

X [Required]

The array of observation for the variable in columns.

S [Optional]

If Scale = G, S must be present and contain the scaling factors for variables.

Input_Dist_Mat [Optional]

Contains the strictly lower triangle of the distance matrix to be updated, stored packed by rows, i.e., in the same format as produced by DIST_MAT. If omitted, no updating takes place.

Output

Output as a **NAGExtract format list**.

Distance Matrix

Note:

The matrix is returned as a lower triangular matrix.

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.3.3. CLUSTER (Set 1)

CLUSTER (NAG G03ECF) performs hierarchical cluster analysis.

Parameters**Method** [Required]

Indicates which clustering method is used.

- 1: single link.
- 2: complete link.
- 3: group average.
- 4: centroid.
- 5: median.
- 6: minimum variance.

Note: EITHER the Distance_matrix OR X must be supplied.

Distance_matrix [Optional]

The strictly lower triangular distance matrix. This can either be packed by rows as output directly by DIST_MAT or as the lower triangular part of a two dimensional array with $n-1$ rows and $n-1$ columns, i.e., without the diagonal cells. Negative distances are not allowed.

X [Optional]

The array of observation for the variable in columns

Distance [Optional]

Should only be present if X is present, Distance indicates which type of distances are computed.

- A: absolute distances.
- E: Euclidean distances.
- S: Euclidean Squared distances.

Scaling [Optional]

Should only be present if X is present, Scaling indicates the standardization of the variables to be used.

- S: standard deviation.
- R: range.
- U: unscaled.

Names [Optional]

The names of the objects to be clustered.

Default: Obj 1, Obj 2, etc.

Output

Output as a **NAGExtract** format list.

Cluster Steps**Dendrogram information**

Notes:

The Cluster steps array consists of four columns. If there are n objects then there are $n-1$ steps in the clustering process. At a step two, clusters are fused in which one cluster (called the second cluster) is fused into the first cluster. The name of the cluster is then always the name of the first object to enter the cluster in its evolution. For each of the $n-1$ clustering steps the following is returned.

- Column 1: The step number
- Column 2: Name of first cluster
- Column 3: Name of second cluster
- Column 4: Distance

The Dendrogram information array consists of three columns. The second column is the name of the objects in the order that is required for a dendrogram and the third is the height associated with that branch of the dendrogram. The dendrogram can be constructed by drawing lines of the required height corresponding to each object in the dendrogram order and then connecting the top of each line with the next line to the right at that level. Finally, the first column is an index that relates dendrogram order to the order of the object in the original data.

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.3.4. GROUPS_FROM_CLUSTER (Set 1)

GROUPS_FROM_CLUSTER (NAG G03EJF) computes a cluster indicator variable from the results of a cluster analysis performed by CLUSTER.

Parameters**Objects** [Required]

The object names in dendrogram order as output by CLUSTER in the second column of Dendrogram information.

Distances [Required]

The distances as output by CLUSTER in the second column Dendrogram information.

Note: EITHER No_of_clusters OR Cluster_distance must be provided.

No_of_clusters [Optional]

The number of clusters required.

Cluster_distance [Optional]

The distance at which clusters are to be produced.

Index [Optional]

If present, the output cluster indicator will be in the order of the original data rather than the dendrogram order. Index should be the first column of the Dendrogram information produced by CLUSTER.

Output

Output as a **NAGExtract** format list.

Number of Clusters**Cluster Distance****Group Index****Note:**

The group index will be for objects in dendrogram order unless Index is provided, in which case it will be in the order of the original data.

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.3.5. DISCRIM_ALLOC (Set 1)

DISCRIM_ALLOC (NAG G03DCF) allocates observations to groups according to selected rules. It is intended for use after DISCRIM_TEST.

Parameters**Info** [Required]

The information provided by DISCRIM_TEST. Either the complete output should be input or DISCRIM_TEST should be used as the argument.

X [Required]

The array of observations to be allocated using the information in Info. X, along with Select_x if present, must be compatible with the input to DISCRIM_TEST.

Select_x [Optional]

If $\text{Select_x}(i) > 0$ then the variable in the i th column of the array X is included in the analysis otherwise it is excluded. This should be compatible with the input to DISCRIM_TEST.

Default: all columns are included.

Equal [Optional]

Indicates whether or not the within-group variance-covariance matrices are assumed to be equal and the pooled variance-covariance matrix used.

- E: the within-group variance-covariance matrices are assumed to be equal.
- U: the within-group variance-covariance matrices are assumed to be unequal.

Default: the within-group variance-covariance matrices are assumed to be unequal.

Approach [Optional]

Indicates which approach is to be used.

- E: the estimative approach is used.
- P: the predictive approach is used.

Default: the predictive approach is used.

Priors [Optional]

The prior probabilities for allocation to the groups.

Default: equal prior probabilities.

Output

Output as a **NAGExtract format list**

Allocated group
Posterior probabilities
Atypicality index

Note:

Both the posterior probabilities and the atypicality index return one column for each group.

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.3.6. DISCRIM_DIST (Set 1)

DISCRIM_DIST (NAG G03DBF) computes Mahalanobis squared distances for group or pooled variance-covariance matrices. It is intended for use after DISCRIM_TEST.

Parameters**Info** [Required]

The information provided by DISCRIM_TEST. Either the complete output should be input or DISCRIM_TEST should be used as the argument.

Equal [Required]

Indicates whether or not the within-group variance-covariance matrices are assumed to be equal and the pooled variance-covariance matrix used.

- E: the within-group variance-covariance matrices are assumed to be equal.
- U: the within-group variance-covariance matrices are assumed to be unequal.

X [Optional]

If present, the distances from the group means to the points in the array X will be calculated, otherwise the distances between the group means will be calculated. X, along with Select_x if present, must be compatible with the input to DISCRIM_TEST.

Default: between group mean distances calculated.

Select_x [Optional]

Should only be present if X is present. If $\text{Select_x}(i) > 0$ then the variable in the i th column of the array X is included in the analysis otherwise it is excluded. This should be compatible with the input to DISCRIM_TEST.

Default: all columns are included.

Output

Output as a **NAGExtract format list**

Distances

Note:

If X is supplied then the distances are an array with rows for the observations and columns for the groups. Otherwise, if Equal = U, the between-group mean distances are in a square array with blank diagonal cells, the within group covariance matrix for the column group being used to standardize the distances. If Equal = E then only the lower triangle of symmetric between group distances is returned.

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.3.7. DISCRIM_TEST (Set 1)

DISCRIM_TEST (NAG G03DAF) computes a test statistic for the equality of within-group covariance matrices and also computes information for use in discriminant analysis.

Parameters**In_group** [Required]

The column that indicates to which group the observations. Note that each group must have at least as many members as there are variables in the analysis.

X [Required]

The array containing variables to be used in the analysis, in columns.

Select_x [Optional]

If $\text{Select_x}(i) > 0$ then the variable in the i th column of the array X is included in the analysis otherwise it is excluded.

Default: all columns included.

Weights [Optional]

If required, the weights for each the observations. If $\text{Weights}(i) = 0.0$ then the i th observation is not included in the analysis. If present, $\text{Weights} \geq 0.0$.

Output

Output as a **NAGExtract format list**

[Test statistic](#)
[Degrees of freedom](#)
[Significance](#)
[Discrim. Info](#)
[Number in group](#)
[Group means](#)

Note:

Since DISCRIM_TEST is intended for use before DISCRIM_ALLOC and DISCRIM_DIST, the output contains information required by these functions. This is held in Discrim. Info.

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.3.8. MDS (Set 1)

MDS (NAG G03FCF) performs non-metric (ordinal) multidimensional scaling.

Parameters

Distance_matrix [Required]

The array containing strictly lower triangle of the distance matrix, i.e., no diagonal elements. Note that negative distances are assumed to represent missing values.

Initial_x [Required]

The array of the initial estimates of the coordinates for point, one column per dimension required. Often the results from METRIC_SCAL can be used as suitable initial estimates.

STRESS_type [Optional]

Indicates whether STRESS or Squared STRESS is to be used as the criterion.

T: STRESS is used.

S: squared STRESS is used.

Max_iter [Optional]

The maximum number of iterations to be used in the fitting process. If present, Max_iter > 0.

Default: 50

Output_distns [Optional]

This optional parameter may be used to select the distances returned:

Output_distns = 0 no distances returned

Output_distns = 1 fitted distances returned

Output_distns > 1 four sets of distances are returned:

Fitted distances

Fitted distances ordered by input distances

Monotonic distances ordered by input distances

Monotonic distances

Default: 0 – no distances returned.

Output

Output as a **NAGExtract** format list.

Coordinates

STRESS

Fitted distances

Ordered fitted distances

Ordered monotonic distances

Monotonic distances

Note:

The fitted distances and monotonic distances are returned in lower triangular form corresponding to the input distance matrix while the Ordered fitted distances and Ordered monotonic distances are returned as one single column.

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

5.5.3.9. METRIC_SCAL (Set 1)

METRIC_SCAL (NAG G03FAF) performs a principal coordinate analysis also known as classical metric scaling.

Parameters

Distance_matrix [Required]

The array containing strictly lower triangle of the distance matrix, i.e., no diagonal elements. Note that negative distances are not allowed.

Dimensions [Optional]

The number of dimensions used to represent the data.

Default: 2

Roots [Optional]

Indicates if all the eigenvalues are to be computed or just the largest corresponding to the number of dimensions.

- A: all the eigenvalues are computed.
- L: only the largest eigenvalues are computed.

Default: L – Largest only.

Output

Output as a **NAGExtract format list**.

Coordinates

Eigenvalues

Remarks

See 'Using the Add-Ins' and the Introduction to Multivariate Methods for further information.

6. NAG Books Set 2 (Nonparametric Methods)

6.1. *Introduction*

There are 3 Add-Ins provided with the NAG Statistical Add-Ins for Excel: Nonparametric Methods product:

NAG Add-In Utilities – A set of utilities used by the other NAG Books

Book 4 – Nonparametric statistics

NAGExtract – A utility to reposition output more conveniently on the worksheet

6.2. *Book 4*

6.2.1. Introduction to Nonparametric Methods

Nonparametric statistics have two advantages over distribution based statistics. The first advantage is that they do not rely on the form of the distribution of the data. A standard assumption is that the data has a Normal distribution. Based on this assumption, various statistical tests have been derived. However, there are many circumstances in which the assumption of Normally distributed data may not be valid, and for small samples it is very difficult to validate the assumption of Normally distributed data with any degree of confidence. The second advantage is that the data may not be actual measurements but are either in the form of classes or in terms of relative comparisons, i.e., ranks. Nonparametric tests generally only require the number in a particular category or the ranks for a set of data.

In order to consider this second point further it is useful to examine a fourfold categorisation of data: nominal (categorical), ordinal, interval, and ratio.

1. The nominal scale is used only to categorise data; for each category a name, perhaps numeric, is assigned so that two different categories will be identified by distinct names, for example red, green and blue.
2. The ordinal scale, as well as categorising the observations, orders the categories. Each is assigned a distinct identifying symbol in such a way that the order of the symbols corresponds to the order of the categories, for example small, medium, and large. (The most common system for ordinal variables is to assign numerical identifiers to the categories, though if they have previously been assigned alphabetic characters, these may be transformed to a numerical system by any convenient method which preserves the ordering of the categories.)
3. The interval scale not only categorises and orders the observations, but also quantifies the comparison between categories; this necessitates a common unit of measurement and an arbitrary zero-point, for example temperature in degrees Celsius.
4. The ratio scale is similar to the interval scale, except that it has an absolute (as opposed to arbitrary) zero-point, for example weight in grams.

Nonparametric statistics are conveniently divided into those that operate on nominal data and those that require ordinal data. Clearly, any test can be used on data of a higher level but there may be some loss of information.

Nonparametric statistics can also be categorised by their functions as follows:

1. One sample
2. Matched pairs samples
3. Two samples
4. K samples
5. Correlation

For the first four categories the most common functions available are for statistical tests. In particular for ordinal data the nonparametric tests can be seen as alternatives to the more usual t and F tests for Normally distributed data. The rank based correlation methods provide an alternative to the product moment correlation coefficient.

For nominal data there are two situations. One is when there are only two categories, binary data, in which case the data can be considered to come from a binomial distribution. In addition to statistics that explicitly use the binomial distribution, specialist tests also exist for this type of data, for example the McNemar test for paired data and the Cochran Q test for k related samples. For data with more than two categories, most of the statistics are based on the χ^2 test for contingency tables. For more complex modelling of binary and general nominal data, generalised linear models can be used, in particular the logistic regression model and the log-linear model. Both of these models are available in Book 2 of the NAG Statistical Add-Ins for Excel.

In addition to tests it is also possible to provide interval estimates (confidence intervals) for location parameters based on ranks rather than the t distribution. These provide a robust alternative to the standard methods.

The table below gives a list of the nonparametric functions available in this Book. The first row is for nominal data, the second for ordinal data.

You are advised to consult a standard text book on nonparametric statistics for further information on these tests. In particular, see Siegel (1956) and Conover (1980). Both are available in later editions. Most of the examples given in the Example Workbook are from Siegel.

Table of available functions

Sample type	Nominal	Ordinal
One sample	Binomial test (BINOMIAL_TEST) Chi-squared test for one sample (CHI_SQ_1) Proportion confidence interval (PROPORTION_CI)	Cox–Stuart test (COX_STUART) One sample confidence interval (ONE_SAMPLE_CI) Runs test (RUNS_TEST) Wilcoxon one sample test (WILCOXON)
Matched pairs samples	McNemar test (MCNEMAR)	Sign test (SIGN_PAIRS) Wilcoxon matched pairs test (WILCOXON_PAIRS)
Two samples	Chi-squared test for 2 samples (CHI_SQ_2)	Kolmogorov–Smirnov two sample test (KOLMOGOROV_2) Mann–Whitney test (MANN_WHITNEY) Median test (MEDIAN_TEST) Two sample confidence interval (TWO_SAMPLE_CI)
K samples	Cochran Q test (COCHRAN_Q) Chi-squared test for k samples (CHI_SQ_K)	Friedman test (FRIEDMAN) Kruskal–Wallis test (KRUSKAL_WALLIS)
Correlation and association	Contingency coefficient (CONTINGENCY_C)	Kendall's Coefficient of Concordance (KENDALL_CC) Kendall's tau rank correlation (KENDALL_TAU) Spearman's rho rank correlation (SPEARMAN_RHO)

An online version of the table is available using the function NONPARAMETRIC_INDEX.

6.2.2. Index of Nonparametric Methods

6.2.2.1. NONPARAMETRIC_INDEX (Set 2)

NONPARAMETRIC_INDEX returns a list of the nonparametric functions available in the NAG Add-Ins.

Parameters

Class [Optional]

Indicates which class of functions is to be returned.

- A: all functions.
- O: one sample tests and confidence intervals.
- M: matched pairs tests.
- T: two sample tests and confidence intervals.

K: K sample tests.
C: rank correlation coefficients.

Default: all functions.

Data_type [Optional]

Indicates for which type of data the functions returned should be suitable.

A: all functions.
O: ordinal.
N: nominal.

Default: all functions.

Output

Output as a **NAGExtract** format list.

List of functions

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.3. One Sample

6.2.3.1. BINOMIAL_TEST (Set 2)

BINOMIAL_TEST (NAG G01BJF, G01EAF) computes the binomial test for the value of a proportion estimated from one sample.

Parameters

K [Required]
The observed number of 'successes'.

N [Required]
The number of observations.

P0 [Required]
The proportion to be tested, i.e., the value (p_0) of p under H_0 .

Tail [Optional]
Indicates the critical region and hence the form of H_1 .

T: two tailed test, i.e., $H_1 = p \neq p_0$
L: lower tailed test, i.e., $H_1 = p < p_0$
U: upper tailed test, i.e., $H_1 = p > p_0$

Default: two tailed test.

Output

Output as a **NAGExtract** format list.

Z value, i.e., normalised test statistic
Significance level

The significance level is exact for $N < 20$ and uses a Normal approximation for larger numbers of observations.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.3.2. CHI_SQ_1 (Set 2)

CHI_SQ_1 (NAG G01ECF) computes the chi-squared test for one sample.

Parameters

Observed [Required]

The observed frequencies for each class.

Expected [Required]

The expected values for each class. These can be either probabilities or expected frequencies. The values will be scaled to sum to the total number observations.

Estim_param [Optional]

The number of estimated parameters used to calculate the fitted values.

Default: 0.

Output

Output as a **NAGExtract** format list.

Chi-squared value

Chi-squared significance level

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.3.3. COX_STUART (Set 2)

COX_STUART (NAG G08AAF) computes the Cox–Stuart test for trend with ordinal from one sample.

Parameters

Series [Required]

The observed series.

Tail [Optional]

Indicates the critical region and hence the form of H_1 .

T: two tailed test, i.e., test for increasing or decreasing trend.

L: lower tailed test, i.e., test for decreasing trend.

U: upper tailed test, i.e., test for increasing trend.

Default: two tailed test.

Output

Output as a **NAGExtract** format list.

Rank sum

Significance level

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information

6.2.3.4. RUNS_TEST (Set 2)

RUNS_TEST (NAG G01ALF, G01EAF) computes the runs test for a single series of ordinal data.

Parameters

Series [Required]

The observed series.

Cut_Point [Optional]

The cut point that defines the "runs below" (values $\leq$ cut point) or "runs above" (values $>$ cut point).

Default: sample median.

Output

Output as a **NAGExtract format list**.

Cut point

Number of runs

Number of runs below the cut point

Number of runs above the cut point

Test statistic

Significance level

The significance uses a Normal approximation.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods statistics for further information.

6.2.3.5. WILCOXON (Set 2)

WILCOXON (NAG G08AGF) computes the Wilcoxon test for the median for ordinal data from a single sample.

Parameters

Data [Required]

The observations from the sample.

Median [Required]

The value of the median under the null hypothesis, m_0 .

Tail [Optional]

Indicates the critical region and hence the form of H_1 .

T: two tailed test, i.e., $H_1 = m \neq m_0$

L: lower tailed test, i.e., $H_1 = m < m_0$

U: upper tailed test, i.e., $H_1 = m > m_0$

Default: two tailed test.

Zeros [Optional]

Indicates whether or not to include cases when the observation equals m_0 in the calculation of the rank sum.

Y: include zeros.
N: do not include zeros.

Default: zeros are not included.

Output

Output as a **NAGExtract format list**.

Rank sum

Z value, i.e., normalised test statistic

Significance level

Number of non-zero values

The significance level is exact for number of observations < 80 and uses a Normal approximation for larger numbers of observations.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.3.6. ONE_SAMPLE_CI (Set 2)

ONE_SAMPLE_CI (NAG G07EAF) computes a rank based confidence interval for the location of a single sample of ordinal data.

Parameters

Data [Required]

The observations from the sample.

Confidence [Optional]

The required confidence for the interval in the range 0 to 1.

Default: 0.95, i.e., a 95% confidence interval

Method [Optional]

Indicates which algorithm is used to compute the interval.

E: exact algorithm
A: iterative algorithm

Default: exact algorithm.

Output

Output as a **NAGExtract format list**.

Point estimate

Lower limit

Upper limit

Estimated confidence

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.3.7. PROPORTION_CI (Set 2)

PROPORTION_CI (NAG G07AAF) computes a confidence interval for the value of a proportion estimated from a single sample.

Parameters

K [Required]
The observed number of 'successes'.

N [Required]
The number of observations.

Confidence [Optional]
The required confidence for the interval in the range 0 to 1.

Default: 0.95, i.e., a 95% confidence interval.

Output

Output as a **NAGExtract** format list.

Lower limit
Upper limit

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.4. Matched Pairs Samples

6.2.4.1. MCNEMAR (Set 2)

MCNEMAR (NAG G01BJF, G01ECF) computes the McNemar test for two matched samples of nominal (binary) data.

Parameters

TABLE [Required]
A 2 by 2 array of cells containing the counts of observations on the two matched samples. If the observations take the values 0 or 1 then:
Cell (1,1) = number of observations for which sample one is 0 and sample two is 0,
Cell (1,2) = number of observations for which sample one is 0 and sample two is 1,
Cell (2,1) = number of observations for which sample one is 1 and sample two is 0,
Cell (2,2) = number of observations for which sample one is 1 and sample two is 1,

Output

Output as a **NAGExtract** format list.

Z value, i.e., normalised test statistic
Significance level

The significance level is exact for $N < 20$ and uses a Normal approximation for larger numbers of observations.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.4.2. SIGN_PAIRS (Set 2)

SIGN_PAIRS (NAG G08AAF) computes the sign test for two matched samples of ordinal data.

Parameters

X [Required]

The observations for sample one.

Y [Required]

The observations for sample two.

Tail [Optional]

Indicates the critical region and hence the form of H_1 .

- T: two tailed test, i.e., $H_1 = m_X \neq m_Y$
- L: lower tailed test, i.e., $H_1 = m_X < m_Y$
- U: upper tailed test, i.e., $H_1 = m_X > m_Y$

where m_X and m_Y are the locations of the populations sampled in X and Y.

Default: two tailed test.

Output

Output as a **NAGExtract** format list.

Test statistic**Significance level****Number of non-tied pairs****Remarks**

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.4.3. WILCOXON_PAIRS (Set 2)

WILCOXON_PAIRS (NAG G08AGF) computes the Wilcoxon test for two matched samples of ordinal data where the differences are also ordinal.

Parameters

X [Required]

The observations for sample one.

Y [Required]

The observations for sample two.

Tail [Optional]

Indicates the critical region and hence the form of H_1 .

- T: two tailed test, i.e., $H_1 = m_X \neq m_Y$
- L: lower tailed test, i.e., $H_1 = m_X < m_Y$
- U: upper tailed test, i.e., $H_1 = m_X > m_Y$

where m_X and m_Y are the locations of the populations sampled in X and Y.

Default: two tailed test.

Zeros [Optional]

Indicates whether or not to include cases when the observation equals m_0 in the calculation of the rank sum.

Y: include zeros.
N: do not include zeros.

Default: zeros are not included.

Output

Output as a **NAGExtract** format list.

Rank sum

Z value, i.e., normalised test statistic

Significance level

Number of non-zero values

The significance level is exact for number of observations < 80 and uses a Normal approximation for larger numbers of observations.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.5. Two Samples

6.2.5.1. CHI_SQ2 (Set 2)

CHI_SQ_2 (NAG G11AAF) computes the chi-squared test for two independent samples of nominal data.

Parameters

Data [Required]

The n rows by 2 columns array of the observations where n is the number of classes, n must be greater or equal to 2.

Output

Output as a **NAGExtract** format list.

Significance

Chisquared

Degrees of freedom

Expected values

Chisquared contributions

The significance returned is for the chi-squared statistic except for a 2×2 table when the total number of observations is less than or equal to 40, in which case the probability from Fisher's exact test is returned.

The chi-squared contributions are the values of $(\text{observed} - \text{expected})^2 / \text{expected}^2$ for each cell which make up the chi-squared statistic.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.5.2. KOLMOGOROV_2 (Set 2)

KOLMOGOROV_2 (NAG G08CDF) computes the sample Kolmogorov–Smirnov test

Parameters

X [Required]

The values of the first sample.

Y [Required]

The values of the second sample.

Tail [Optional]

Indicates the critical region and hence the form of H_1 .

- T: two tailed test, i.e., distributions are not equal.
- L: lower tailed test, i.e., X values tend to be smaller than Y values.
- U: upper tailed test, i.e., X values tend to be larger than Y values.

Default: two tailed test.

Output

Output as a **NAGExtract** format list.

Test statistic

Z value, i.e., normalised test statistic

Significance level

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.5.3. MANN_WHITNEY (Set 2)

MANN_WHITNEY (NAG G08AHF) computes the Mann–Whitney test for two independent samples of ordinal data.

Parameters

X [Required]

The observations for sample one.

Y [Required]

The observations for sample two.

Tail [Optional]

Indicates the critical region and hence the form of H_1 .

- T: two tailed test, i.e., $H_1 = m_X \neq m_Y$
- L: lower tailed test, i.e., $H_1 = m_X < m_Y$
- U: upper tailed test, i.e., $H_1 = m_X > m_Y$

where m_X and m_Y are the locations of the populations sampled in X and Y.

Default: two tailed test.

Output

Output as a **NAGExtract** format list.

Rank sum

Normalised test statistic

(Exact) Significance level

The significance level is exact for total number of observations < 40 and uses a Normal approximation for larger numbers of observations.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.5.4. MEDIAN_TEST (Set 2)

MEDIAN_TEST (NAG G08ACF) computes the median test for two independent samples of ordinal data.

Parameters

X [Required]

The observations for sample one.

Y [Required]

The observations for sample two.

Output

Output as a **NAGExtract** format list.

Number of observations below the median for sample one

Number of observations below the median for sample two

Significance level

The significance level is exact for number of observation < 40 and uses a Normal approximation for larger numbers of observations.

Remarks

Approximate 1-tail probabilities may be obtained by halving the returned significance level if the difference between the samples in the required direction.

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.5.5. TWO_SAMPLE_CI (Set 2)

TWO_SAMPLE_CI (NAG G07EBF) computes a rank based confidence interval for the difference in location of two samples of ordinal data.

Parameters

X [Required]

The observations from sample one.

Y [Required]

The observations from sample two, the interval is for the location of population two minus the location of population one.

Confidence [Optional]

The required confidence for the interval in the range 0 to 1.

Default: 0.95, i.e., a 95% confidence interval

Method [Optional]

Indicates which algorithm is used to compute the interval.

E: exact algorithm
A: iterative algorithm

Default: exact algorithm.

Output

Output as a **NAGExtract** format list.

Point estimate**Lower limit****Upper limit****Estimated confidence****Remarks**

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.6. K Samples

6.2.6.1. CHI_SQ_K (Set 2)

CHI_SQ_K (NAG G11AAF) computes the chi-squared test for K independent samples of nominal data.

Parameters**Data** [Required]

The n rows by k columns array of the observations, where k is the number of samples being compared and n is the number of classes. n must be greater than 1 and k must be greater than 2.

Output

Output as a **NAGExtract** format list.

Significance**Chisquared****Degrees of freedom****Expected values****Chisquared contributions**

The chi-squared contributions are the values of $(\text{observed} - \text{expected})^2 / \text{expected}^2$ for each cell which make up the chi-squared statistic.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.6.2. COCHRAN_Q (Set 2)

COCHRAN_Q (NAG G08ALF) computes the Cochran Q test for nominal (binary) data from k samples.

Parameters

Data [Required]

The n rows by k columns array of 0 or 1 values, where n is the number of "blocks" and k is the number of "treatments". Both n and k must be greater than 1.

Output

Output as a **NAGExtract format list**.

Test statistic
Significance level

The significance level is based on a chi-squared approximation and should be reasonable if, after omitting rows which contain either all 0s or all 1s there are more than 5 rows and 25 observations.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.6.3. FRIEDMAN (Set 2)

FRIEDMAN (NAG G08AEF) computes the Friedman test for ordinal data from k matched samples.

Parameters

Data [Required]

The n rows by k columns array of the observations, where n is the number of "blocks" and k is the number of "treatments". Both n and k must be greater than 1.

Output

Output as a **NAGExtract format list**.

Test statistic
Significance level

The significance level is based on a chi-squared approximation that is reasonable provided $k = 2$ and $n > 19$ or $k = 3$ and $n > 9$ or $k = 4$ or $n > 4$ or $k > 4$.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.6.4. KRUSKAL_WALLIS (Set 2)

KRUSKAL_WALLIS (NAG G08AFF) computes the Kruskal–Wallis test for k independent samples or groups of ordinal data.

Parameters

Data [Required]

A column containing the observations for all samples.

Groups [Required]

A column containing the indicators as to which group/sample each observation belongs. The indicators can be text or numeric. The number of groups will be calculated from the number of distinct values in Groups.

Output

Output as a **NAGExtract format list**.

Test statistic
Significance level**Remarks**

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.7. Correlation and Association

6.2.7.1. CONTINGENCY_C (Set 2)

CONTINGENCY_C (NAG G11AAF) computes the contingency coefficient for nominal data.

Parameters**Data** [Required]

The n rows by k columns array of the observations, where k is the number of samples being compared and n is the number of classes. n must be greater than 1 and k must be greater or equal to 2.

Output

Output as a **NAGExtract format list**.

Contingency Coefficient**Remarks**

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.7.2. KENDALL_CC (Set 2)

KENDALL_CC (NAG G08DAF) computes Kendall's coefficient of concordance.

Parameters**Data** [Required]

The n rows by k columns array of the observations, where k is the number of objects being compared and n is the number of comparisons. Both n and k must be greater than 1.

Output

Output as a **NAGExtract format list**.

Kendall's coefficient of concordance
Approximate significance level**Remarks**

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.7.3. KENDALL_TAU (Set 2)

KENDALL_TAU (NAG G02BNF) computes the Kendall's rank correlation coefficient, τ .

Parameters**Data** [Required]

The n rows by k columns array of the data, where k is the number of variables and n is the number of observations. Both n and k must be greater than 1.

Output

Output as a **NAGExtract format list**.

Kendall's τ

If there are more than two variables the correlations are returned as a k by k matrix.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

6.2.7.4. SPEARMAN_RHO (Set 2)

SPEARMAN_RHO (NAG G02BNF) computes the Spearman's rank correlation coefficient, r .

Parameters**Data** [Required]

The n rows by k columns array of the data, where k is the number of variables and n is the number of observations. Both n and k must be greater than 1.

Output

Output as a **NAGExtract format list**.

Spearman's r

If there are more than two variables the correlations are returned as a k by k matrix.

Remarks

See 'Using the Add-Ins' and the Introduction to Nonparametric Methods for further information.

7. Troubleshooting Guide

7.1. *Introduction*

This guide is intended to address some of the common problems that users have experienced in the past. This revision has tried to remove some of the difficulties and hopefully this manual will also help. If, nevertheless, your problem is not listed please do not hesitate to contact us at one of the addresses given below and we will try to solve your problem. It will also tell us where the manual or software needs improvement, so we welcome your call.

Please note that complete sets of examples are provided in the workbooks Book0E.xls, Book1E.xls, Book2E.xls, Book3E.xls for the NAG Statistical Add-Ins for Excel: Modelling and Multivariate Methods product in and Book4E.xls for the companion product providing nonparametric statistical capability. These are installed in the 'Examples' subdirectory of the Add-Ins directory and are also on the CD. You might find it helpful to work through these to familiarize yourself with the use of the Add-Ins or to clarify any items not described in sufficient detail.

7.2. *Questions*

What versions of Excel do the Add-Ins support?

The Add-Ins were formed under Excel 7.0 using Windows 95. They have been tested and shown to work under Excel 97 and using Windows NT and Windows 98. We have tested them with Office 2000.

They will not work under Excel 5.0 operating under Windows 95 or Windows NT. If you are interested in a version for this environment, please register an interest with NAG. Should there be sufficient demand then we would consider doing a specific implementation.

The first release placed the Add-Ins in a NAG Category; where are they now?

We have not discovered a 'clean' way of producing additional categories within Excel. We created a category by making a dummy Excel 4 macro. This opens up the possibility of creating a new category and we exploited this to create the 'NAG Books' category. The scheme worked most of the time, but on some systems it seemed to fail. Because of this we have reverted to placing the Add-Ins in the 'Statistical' category which already exists.

We get Microsoft Help when we want NAG Help!

NAG help may be obtained from the 'Nag' menu by accessing NAGHelp.

We only obtain a <list of n> heading not a set of numbers

If you are using a NAG function that you are expecting to return several values then the call to this function must be entered as an Excel ARRAY FUNCTION. In practice this means that the destination cells must be pre-selected and that once the formula has been formed in the **formula bar** CONTROL-SHIFT-ENTER must be pressed to enter the formula. You will note that pressing this combination of keys causes the formula to have curly brackets around it in the formula bar. You will NOT get the same effect by explicitly typing these brackets.

Will NAG be producing any more Add-Ins?

We are keen to learn of user requirements in this area. The Add-Ins have been well-received and we now have two distinct sets of statistical Add-Ins: 'Modelling and Multivariate Methods' as one set and 'Nonparametric Methods' as the second set. Both are documented together, though they are sold as separate products. Other Add-Ins are contemplated. Please let us know what you want.

We get error messages whenever we open a second workbook!

Set the properties of the NAG_Library.xla to read only. To do this, find NAG_Library.xla using Windows Explorer and right click on the file. Choose the 'Properties' option and then the 'general' tab. Check the Read-only box.

We get error messages when we load the Add-Ins!

On some systems error messages appear when the Add-Ins are loaded. Try NOT explicitly loading the NAG Add-In utilities, despite the message. In theory it should not have been necessary to explicitly load these since they are '**referenced**' when each of the Add-Ins are formed. In practice, however, we found that several machine-operating-system and Excel-version combinations seemed to require explicit loading. At least one site has experienced a problem when explicitly loading the utilities Add-In, which they cured by simply relying on the referencing mechanism.

Are there any limitations on the size of data sets we can use?

Most of the NAG Statistical Add-Ins for Excel are implemented as **array functions**. This makes it possible for results to be automatically updated on a worksheet when any of the data items change. However there is a Microsoft limit, in excess of 5000, on the number of elements that may be exported from an array function. It follows therefore that, for those routines where the amount of data reflects the amount of output, there will be a practical limit on the volume of data that can be handled correctly. Otherwise, as far as we are aware, there is no direct restriction on the size of the data sets that may be entered from Excel.

The Add-Ins will not load!

This is usually indicative of a faulty installation. Re-install the Add-Ins from the supplied media. Note that it is not sufficient to copy materials from one machine to another. See 'Installation'.

Have you any examples of use?

Please see the workbooks Book0E.xls, Book1E.xls, Book2E.xls, Book3E.xls or Book4E.xls as appropriate. These are supplied with the relevant Add-Ins product.

Can I obtain more copies of the manual?

These are available from NAG Ltd, please see 'Addresses'. Alternatively, the manual is available in Acrobat format on the CD, on the NAG Web sites and of course via the Windows Help system.

Will I be notified of any updates or additional materials?

Please register with us to ensure that we can keep you up to date with future developments of this product line. If you have purchased the Add-Ins via a third party then we are especially keen that you should either complete the registration form in the Appendix (see registration), or register via the WWW. (See www.nag.co.uk or www.nag.com.) In addition to the benefits of

being kept up to date with NAG developments this will also expedite our response to any questions you may subsequently have about the product and its usage.

The 'Nag' menu doesn't appear when I load the NAG Add-Ins!

Make sure that all of the NAG Books, including NAGExtract, are not loaded. Now load the 'NAG Add-In Utilities'. From the 'User Defined' category in the function wizard, choose 'ChangeKey'. Call it with KeyName set to 'About' and NewValue set to 0. Repeat the process with KeyName set to 'Help' and NewValue set to 0.

The 'Nag' menu doesn't disappear when I unload the NAG Add-Ins!

Make sure that all of the NAG Books, including NAGExtract, are not loaded. Now load the 'NAG Add-In Utilities'. From the 'User Defined' category in the function wizard, choose 'ChangeKey'. Call it with KeyName set to 'About' and NewValue set to 0. Repeat the process with KeyName set to 'Help' and NewValue set to 0.

The 'Comments' on the examples worksheets seem truncated using Excel 97!

Select the cell containing the comment. From the 'Insert' menu select 'Comment'. Resizing handles now appear on the comment box and these may be used to extend the box appropriately.

We get 'Trial' message boxes appearing when we load the NAG Add-Ins

This message appears when the software you have loaded is designed to work for a limited period only. It is to remind you to contact NAG to obtain a licensed copy before the time period expires. After this date the Add-Ins will not work.

We get a 'Catastrophic failure' message when we use the NAG Add-Ins

This may be because your trial license has expired. Please contact NAG to obtain a licensed copy. If you have a licensed copy then please contact NAG with the details of your problem.

8. Appendices

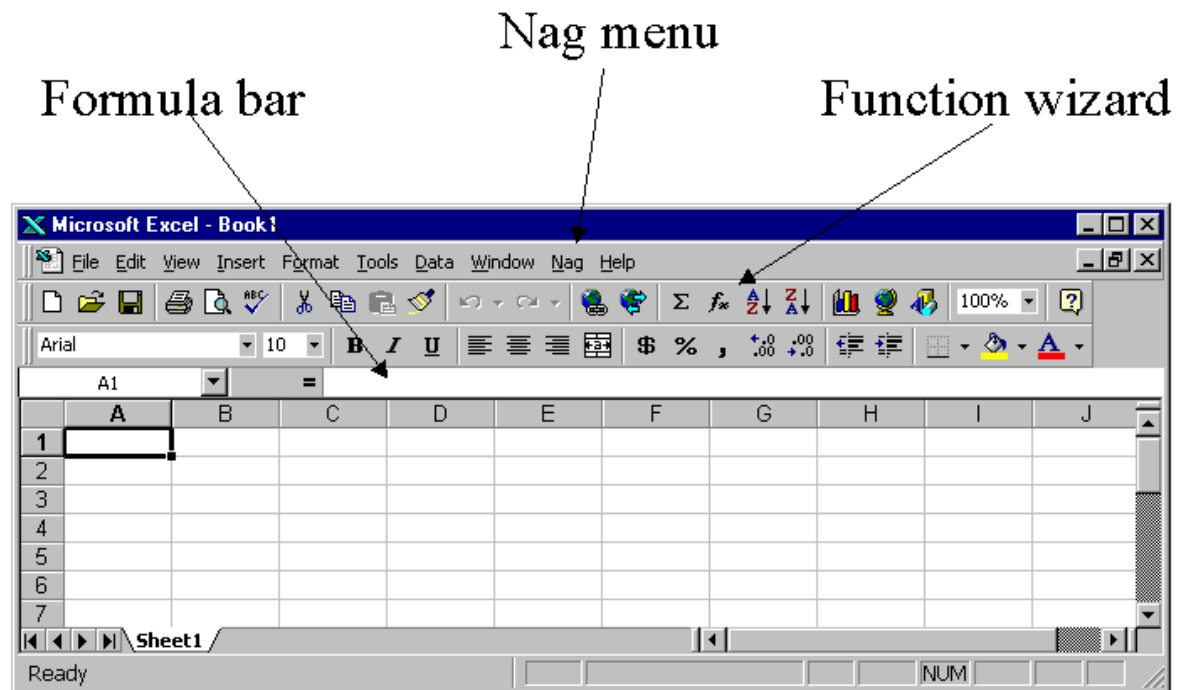
8.1. Glossary

array functions – instead of returning just one scalar value as the result, these functions return more than one value (sometimes as a vector) at other times as arrays. The NAG Statistical Add-In array functions always return results as a column vector. See Array Functions.

function wizard – a dialog box showing the list of available functions and their categories. The wizard then helps you call the selected procedure by providing input boxes for the parameters. See the diagram below.

NAGextract format list – the vector of results returned by any of the NAG Statistical Add-Ins that return more than one scalar value.

reference – a means of getting access to procedures in another workbook.



8.2. *Algorithmic Association*

8.2.1. Add-In / NAG Subroutine Correspondence

Set 1 : Modelling and Multivariate Methods

Available Add-In functions from Book 0:

BETA_PPT	(G01FEF)
BETA_PROB	(G01EEF)
CHI_PPT	(G01FCF)
CHI_PROB	(G01ECF)
F_PPT	(G01FDF)
F_PROB	(G01EDF)
GAMMA_PPT	(G01FFF)
GAMMA_PROB	(G01EFF)
MULTI_NORMAL_PROB	(G01HBF)
NORMAL_PPT	(G01FAF)
NORMAL_PROB	(G01EAF)
S5PT_SUMMARY	(G01ALF)
SUMMARY_STATS	(G01AAF)
T_PPT	(G01FBF)
T_PROB	(G01EBF)
TWO_WAY_TABLE	(G11AAF)

Available Add-In functions from Book 1:

ACF	(G13ABF)
ARIMA_APPROX_FIT	(G13ADF)
ARIMA_FIT	(G13BEF)
ARIMA_FORECAST	(G13AJF)
CORREL_MAT	(G02BXF)
DUMMY_VARS	(G04EAF)
DURBIN_WATSON	(G02FCF)
MULT_LIN_REG	(G02DAF)
PACF	(G13ACF)
PARTIAL_CORREL_MAT	(G02BYF)
SPECTRAL	(G13CBF)
STAND_RESIDS	(G02FAF)
TIME_SERIES_DIFF	(G13AAF)
TRNS_FUNC_FORECAST	(G13BJF)

Available Add-In functions from Book 2:

ANOVA_BLOCK	(G04BBF)
ANOVA_CONTRASTS	(G04DAF)
ANOVA_FACTORIAL	(G04CAF)
ANOVA_ROW_COLUMN	(G04BCF)
BINOMIAL_GLM	(G02GBF)
GAMMA_GLM	(G02GDF)
MULT_COMPN	(G04DBF)
NORMAL_GLM	(G02GAF)
POISSON_GLM	(G02GCF)

Available Add-In functions from Book 3:

CLUSTER	(G03ECF)
DISCRIM_ALLOC	(G03DCF)
DISCRIM_DIST	(G03DBF)
DISCRIM_TEST	(G03DAF)
DIST_MAT	(G03EAF)
FACTOR	(G03CAF)
FACTOR_SCORE	(G03CCF)
GROUPS_FROM_CLUSTER	(G03EJF)
MDS	(G03FCF)
METRIC_SCAL	(G03FAF)
ORTHOG_ROTATION	(G03BAF)
PRIN_COMP	(G03AAF)

Set 2 : Nonparametric Methods

Available Add-In functions from Book 4:

BINOMIAL_TEST	(G01BJF, G01EAF)
CHI_SQ_2	(G11AAF)
CHI_SQ_K	(G11AAF)
CHI_SQ_1	(G01ECF)
COCHRAN_Q	(G08ALF)
CONTINGENCY_C	(G11AAF)
COX_STUART	(G08AAF)
FRIEDMAN	(G08AEF)
KENDALL_CC	(G08DAF)
KENDALL_TAU	(G02BNF)
KOLMOGOROV_2	(G08CDF)
KRUSKAL_WALLIS	(G08AFF)
MANN_WHITNEY	(G08AHF)
MCNEMAR	(G01BJF, G01ECF)
MEDIAN_TEST	(G08ACF)
NONPARAMETRIC_INDEX	
ONE_SAMPLE_CI	(G07EAF)
PROPORTION_CI	(G07AAF)
RUNS_TEST	(G01ALF, G01AEF)
SIGN_PAIRS	(G08AAF)
SPEARMAN_RHO	(G02BNF)
TWO_SAMPLE_CI	(G07EBF)
WILCOXON	(G08AGF)
WILCOXON_PAIRS	(G08AGF)

8.2.2. NAG Subroutine / Add-In Correspondence

Available Add-In functions from Chapter G01:

(G01AAF)	SUMMARY_STATS
(G01AEF)	RUNS_TEST
(G01ALF)	RUNS_TEST
(G01ALF)	S5PT_SUMMARY
(G01BJF)	BINOMIAL_TEST
(G01BJF)	MCNEMAR

(G01EAF)	BINOMIAL_TEST
(G01EAF)	NORMAL_PROB
(G01EBF)	T_PROB
(G01ECF)	CHI_PROB
(G01ECF)	CHI_SQ_1
(G01ECF)	MCNEMAR
(G01EDF)	F_PROB
(G01EEF)	BETA_PROB
(G01EFF)	GAMMA_PROB
(G01FAF)	NORMAL_PPT
(G01FBF)	T_PPT
(G01FCF)	CHI_PPT
(G01FDF)	F_PPT
(G01FEF)	BETA_PPT
(G01FFF)	GAMMA_PPT
(G01HBF)	MULTI_NORMAL_PROB

Available Add-In functions from Chapter G02:

(G02BNF)	KENDALL_TAU
(G02BNF)	SPEARMAN_RHO
(G02BXF)	CORREL_MAT
(G02BYF)	PARTIAL_CORREL_MAT
(G02DAF)	MULT_LIN_REG
(G02FAF)	STAND_RESIDS
(G02FCF)	DURBIN_WATSON
(G02GAF)	NORMAL_GLM
(G02GBF)	BINOMIAL_GLM
(G02GCF)	POISSON_GLM

Available Add-In functions from Chapter G03:

(G03AAF)	PRIN_COMP
(G03BAF)	ORTHOG_ROTATION
(G03CAF)	FACTOR
(G03CCF)	FACTOR_SCORE
(G03DAF)	DISCRIM_TEST
(G03DBF)	DISCRIM_DIST
(G03DCF)	DISCRIM_ALLOC
(G03EAF)	DIST_MAT
(G03ECF)	CLUSTER
(G03EJF)	GROUPS_FROM_CLUSTER
(G03FAF)	METRIC_SCAL
(G03FCF)	MDS

Available Add-In functions from Chapter G04.

(G04BBF)	ANOVA_BLOCK
(G04BCF)	ANOVA_ROW_COLUMN
(G04CAF)	ANOVA_FACTORIAL
(G04DAF)	ANOVA_CONTRASTS
(G04DBF)	MULT_COMPN
(G04EAF)	DUMMY_VARS

Available Add-In functions from Chapter G07.

(G07AAF)	PROPORTION_CI
(G07EAF)	ONE_SAMPLE_CI
(G07EBF)	TWO_SAMPLE_CI

Available Add-In functions from Chapter G08.

(G08AAF)	COX_STUART
(G08AAF)	SIGN_PAIRS
(G08ACF)	MEDIAN_TEST
(G08AEF)	FRIEDMAN
(G08AFF)	KRUSKAL_WALLIS
(G08AGF)	WILCOXON
(G08AGF)	WILCOXON_PAIRS
(G08AHF)	MANN_WHITNEY
(G08ALF)	COCHRAN_Q
(G08CDF)	KOLMOGOROV_2
(G08DAF)	KENDALL_CC

Available Add-In functions from Chapter G11:

(G11AAF)	TWO_WAY_TABLE
(G11AAF)	CHI_SQ_2
(G11AAF)	CHI_SQ_K
(G11AAF)	CONTINGENCY_C

Available Add-In functions from Chapter G13.

(G13AAF)	TIME_SERIES_DIFF
(G13ABF)	ACF
(G13ACF)	PACF
(G13ADF)	ARIMA_APPROX_FIT
(G13AJF)	ARIMA_FORECAST
(G13BEF)	ARIMA_FIT
(G13BJF)	TRNS_FUNC_FORECAST
(G13CBF)	SPECTRAL

8.3. **Registration**

Please register your purchase with us in order to be kept up to date with new revisions or related additional materials. Registration will also enable us to deal more quickly with any subsequent questions you may have about the Add-Ins or their usage.

Simply register via the WWW at www.nag.co.uk or www.nag.com or send a completed copy of this form to one of the NAG addresses.

Name :

Job Title :

Department :

Organization :

Address :

Tel Number :

Fax Number :

Email :

Add-Ins used (e.g. Set 1 and/or Set 2)

Thank you. That is sufficient to register.

If you would be kind enough to complete any or all of the questions below it will help us to continue to enhance and improve the quality products and services that we provide.

1) How easy did you find it to order the software? *Please tick as appropriate.*

Very Easy ☐

Easy ☐

Quite Difficult ☐

Very Difficult ☐

If 'Quite Difficult' or 'Very Difficult' please specify why.

2) How easy did you find the installation of NAG Statistical Add-Ins for Excel? *Please tick as appropriate.*

Very Easy ☐

Easy ☐

Quite Difficult ☐

Very Difficult ☐

If 'Quite Difficult' or 'Very Difficult', please specify why.

3) How useful do you find the Windows Help for the NAG Statistical Add-Ins? *Please tick as appropriate.*

☐ Very Useful ☐ Quite Useful ☐ Of Some Use ☐ Not Useful

If 'Of Some Use' or 'Not Useful', please specify why.

4) Do you like a printed manual? *Please tick as appropriate.*

YES ☐ NO ☐

5) How do you use the NAG Statistical Add-ins? *Please specify.*

6) Did the NAG Statistical Add-Ins replace an existing package? *Please tick as appropriate.*

YES ☐ NO ☐

If YES, please specify what.

7) Do the NAG Statistical Add-Ins meet your needs? *Please tick as appropriate.*

YES ☐ NO ☐

If NO, please specify why.

8) Would you be interested in NAG Statistical Add-Ins that covered the following areas? *Please tick as appropriate.*

	YES	NO
Tabulation	☐	☐
Nonparametric tests	☐	☐
Distribution tests	☐	☐
Survival analysis	☐	☐
Further regression facilities	☐	☐
Handling larger data sets	☐	☐

9) What other enhancements or additions would be of interest to you? *Please specify.*

9. Contacting NAG

NAG Ltd
Wilkinson House
Jordan Hill Road
OXFORD OX2 8DR
United Kingdom

Tel: +44 (0)1865 511245
Fax: +44 (0)1865 310139
Web: www.nag.co.uk

NAG Ltd Response Centre
Email: infodesk@nag.co.uk

Tel: +44 (0)1865 311744
Fax: +44 (0)1865 311755

NAG Inc
1400 Opus Place, Suite 200
DOWNS GROVE, IL 60515-5702
USA

Tel: +1 630 971 2337
Fax: +1 630 971 2706
Web: www.nag.com

NAG Inc Response Center
Email: infodesk@nag.com

Tel: +1 630 971 2345
Fax: +1 630 971 2346

NAG GmbH
Schleissheimer Strasse 5
85748 GARCHING
Deutschland

Tel: +49 (0)89 3207395
Fax: +49 (0)89 3207396
Email: info@naggmbh.de
Web: www.naggmbh.de

Nihon NAG KK
Yaesu Nagaoka Building No 6
1-9-8 Minato
Chuo-ku
TOKYO
Japan

Tel: +81 (0)3 5542 6311
Fax: +81 (0)3 5542 6312
Email: sales@nag-j.co.jp
Web: www.nag-j.co.jp

10. References

- Aitchinson J and Dunsmore I R (1975) *Statistical Prediction Analysis* Cambridge
- Andrews D F and Herzberg A M (1985) *Data : A Collection of Problems from Many Fields for the Student and Research Worker* Springer-Verlag
- Atkinson A C (1981) Two graphical displays for outlying and influential observations in regression *Biometrika* 68 13-20
- Chatfield C (1976) *Statistics for Technology* Chapman and Hall
- Chatfield C (1982) *The Analysis of Time Series: an Introduction* Chapman and Hall
- Cochran W G and Cox G M (1957) *Experimental Designs* Wiley
- Conover W J (1980) *Practical Nonparametric Statistics* Wiley
- Cook R D and Weisberg S (1982) *Residuals and Influence in Regression* Chapman and Hall
- Cox D R (1970) *Analysis of Binary Data* Chapman and Hall
- Davis O L (1978) *The Design and Analysis of Industrial Experiments* Longman
- Dobson A J (1990) *An Introduction to Generalized Linear Models* Chapman and Hall
- Draper N R and Smith H (1966) *Applied Regression Analysis* Wiley
- Everitt B S (1974) *Cluster Analysis* Heinemann
- Everitt B S (1977) *The Analysis of Contingency Tables* Chapman and Hall
- Hand D J, Daly F, Lunn A D, McConway K J and Ostrowski E (1994) *Small Data Sets* Chapman and Hall
- John J A (1987) *Cyclic Designs* Chapman and Hall
- John J A and Quenouille M H (1977) *Experiments: Design and Analysis* Griffin
- Krzanowski W J (1990) *Principles of Multivariate Analysis* Oxford University Press
- Lawley D N and Maxwell A E (1971) *Factor Analysis as a Statistical Method* Butterworths
- McCullagh P and Nelder J A (1989) *Generalized Linear Models* Chapman and Hall
- Morrison D F (1967) *Multivariate Statistical Methods* McGraw-Hill
- Osborn J F (1979) *Statistical Exercises in Medical Research* Blackwell Scientific Publications
- Searle S R (1971) *Linear Models* Wiley
- Siegel S (1956) *Nonparametric Statistics for the Behavioral Sciences* McGraw-Hill
- Snedecor G W and Cochran W G (1967) *Statistical Methods* Iowa State University Press
- Wetherill G B (1972) *Elementary Statistical Methods* Chapman and Hall
- Winer B J (1970) *Statistical Principles in Experimental Design* McGraw-Hill